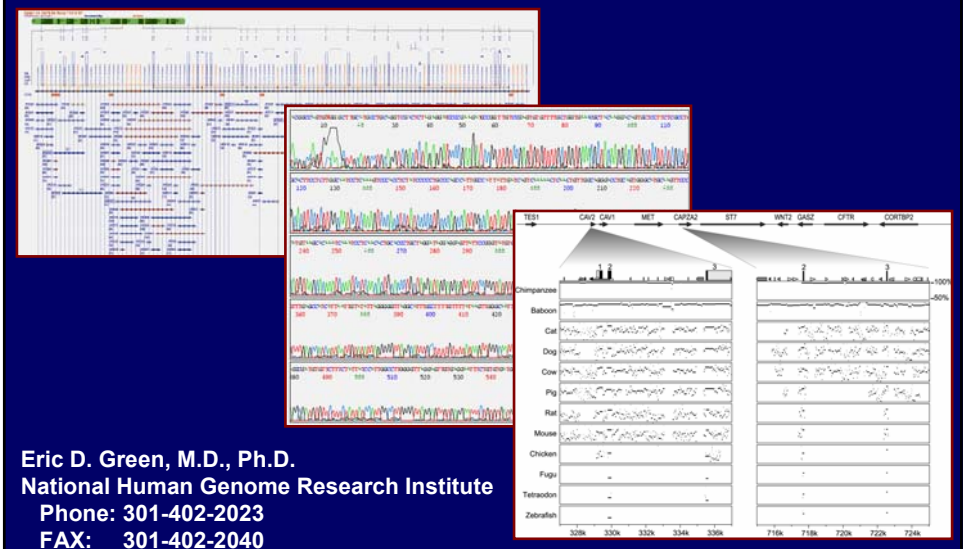
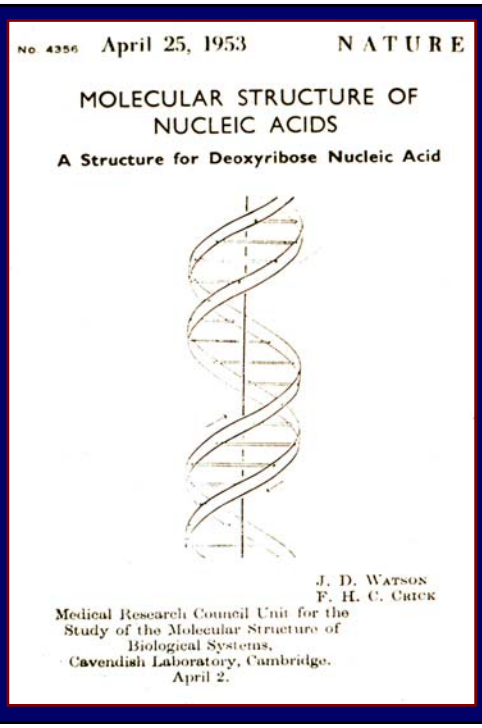
Techniques for Genome Mapping & Sequencing



Eric D. Green, M.D., Ph.D.
National Human Genome Research Institute
Phone: 301-402-2023
FAX: 301-402-2040
E-Mail: egreen@nhgri.nih.gov



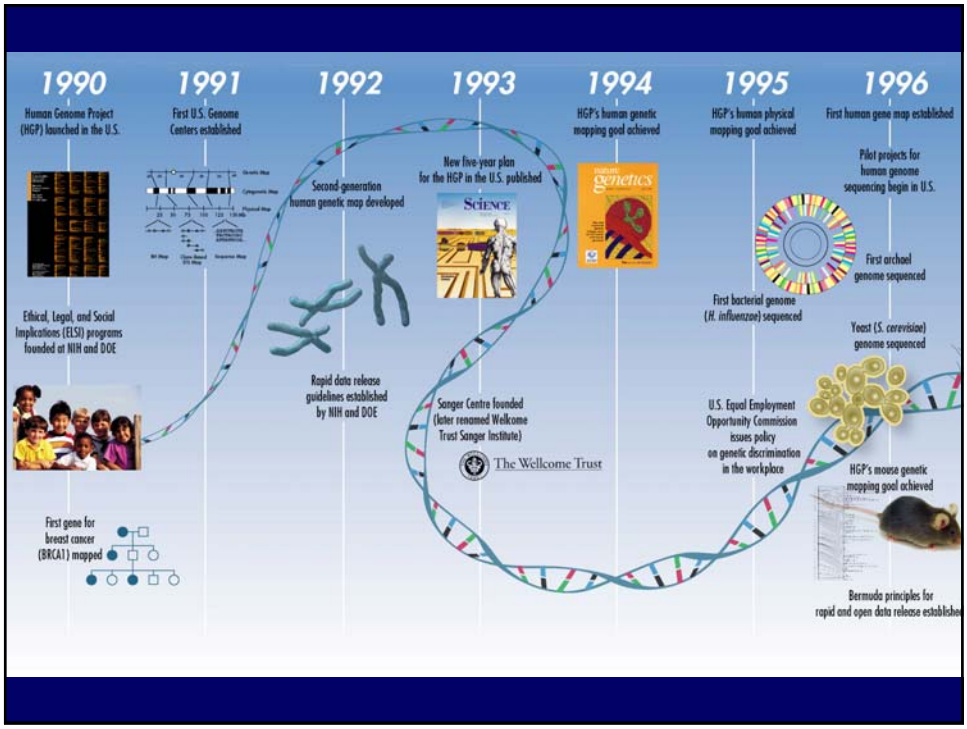
No. 4356 April 25, 1953 NATURE

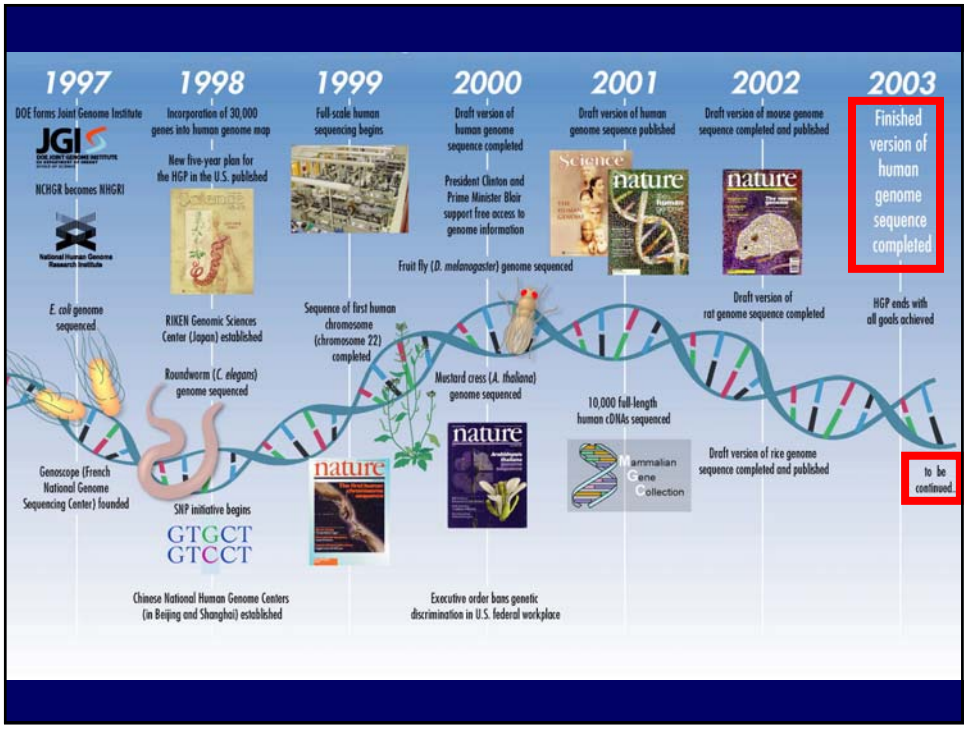
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

J. D. WATSON
F. H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.


1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
Science
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
nature genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release establishe


1997
DOE forms Joint Genome Institute
JGI
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
Science
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
nature
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
nature
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued...

Outline

I. Fundamentals of Genome Mapping

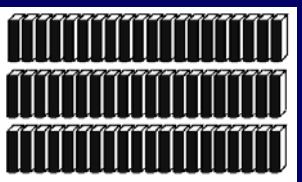
II. Fundamentals of Genome Sequencing

III. Mapping & Sequencing in the Human Genome Project… and Beyond

IV. Comparative Sequencing

Genome Sizes

Human Genome
Mouse Genome



~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

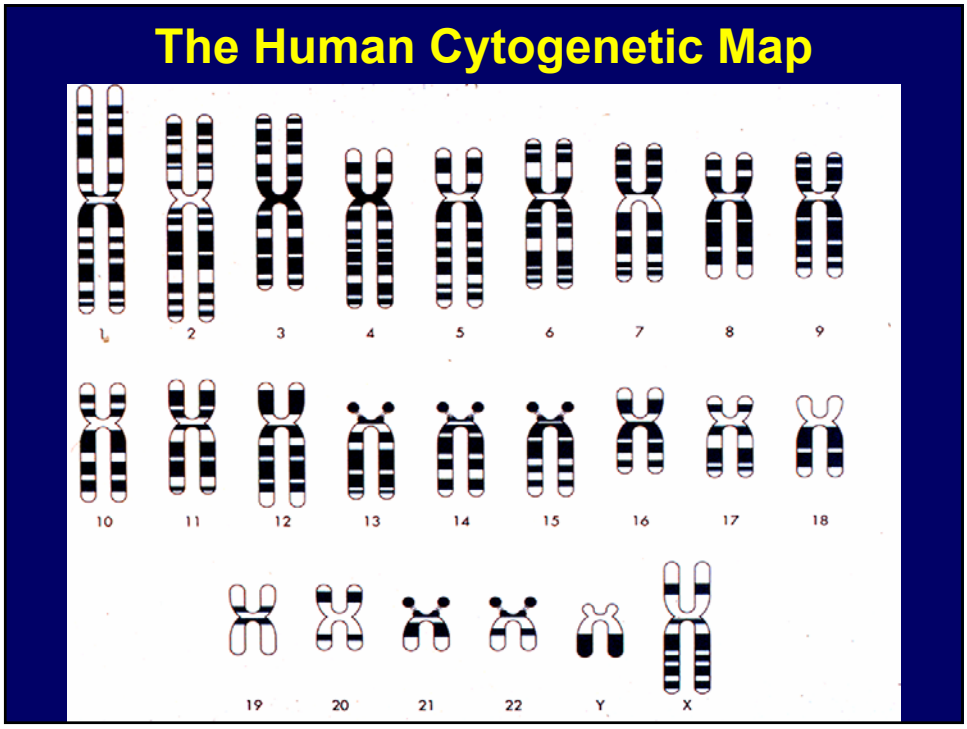
~100,000,000 bp

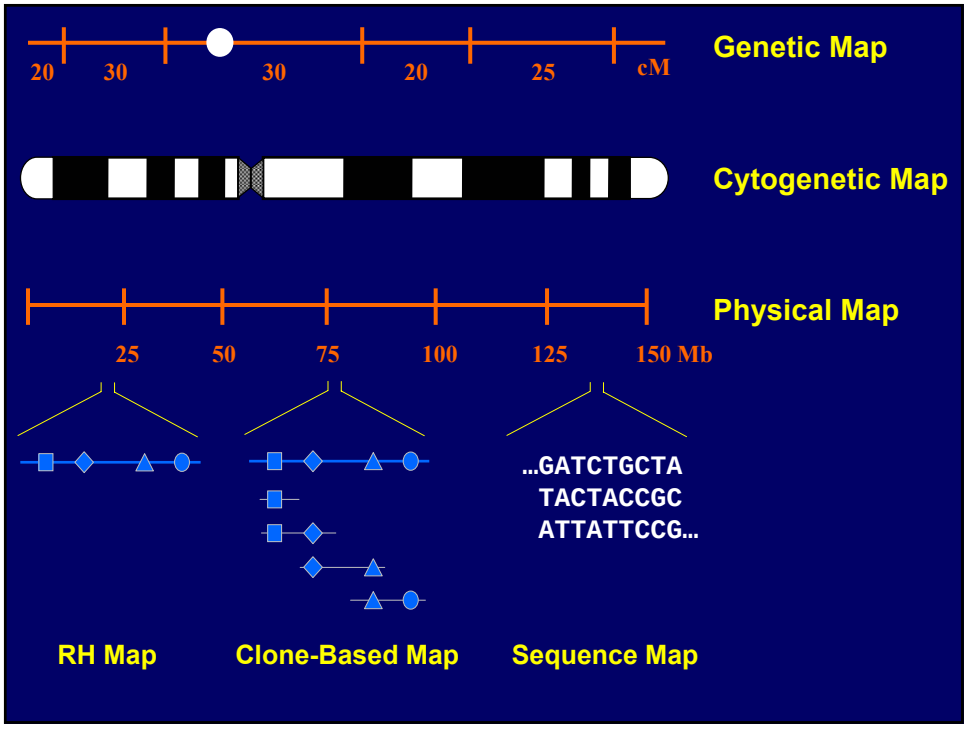
Yeast Genome

~15,000,000 bp

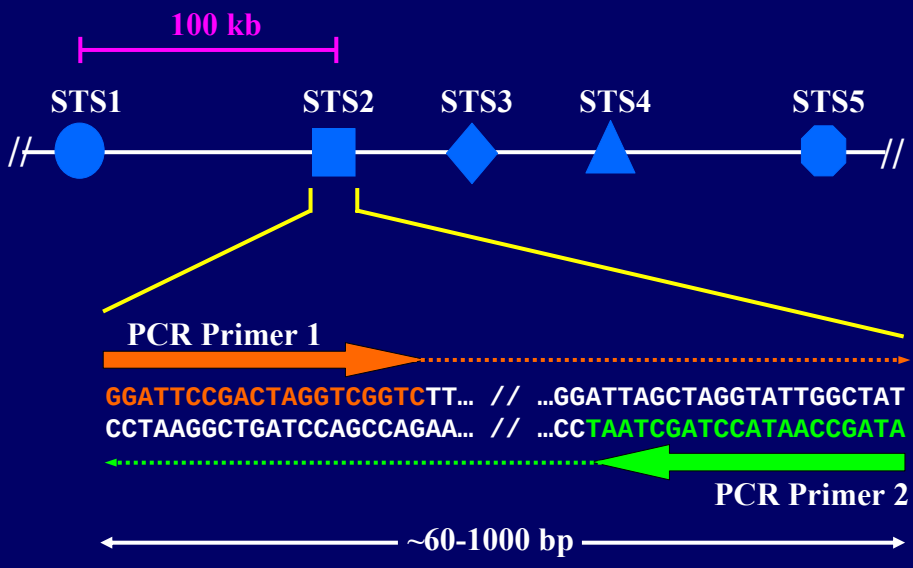
E. coli Genome

~5,000,000 bp


The Human Cytogenetic Map
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Y
X

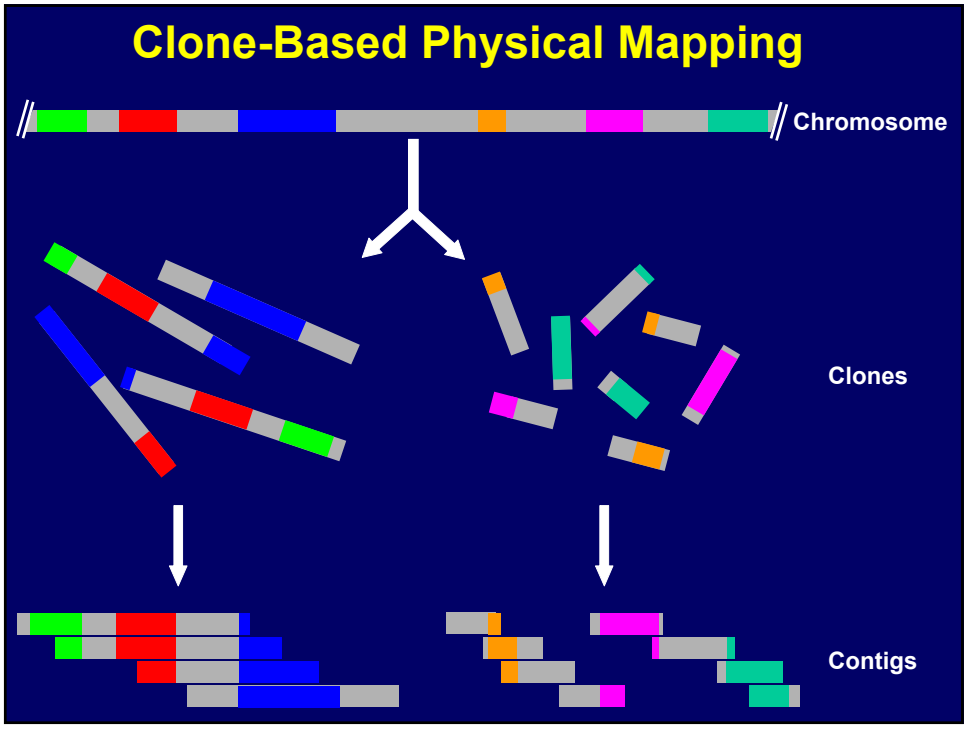

20
30
30
20
25
cM
Genetic Map
Cytogenetic Map
25
50
75
100
125
150 Mb
Physical Map
...GATCTGCTA
TACTACCGC
ATTATTCCG...
RH Map
Clone-Based Map
Sequence Map

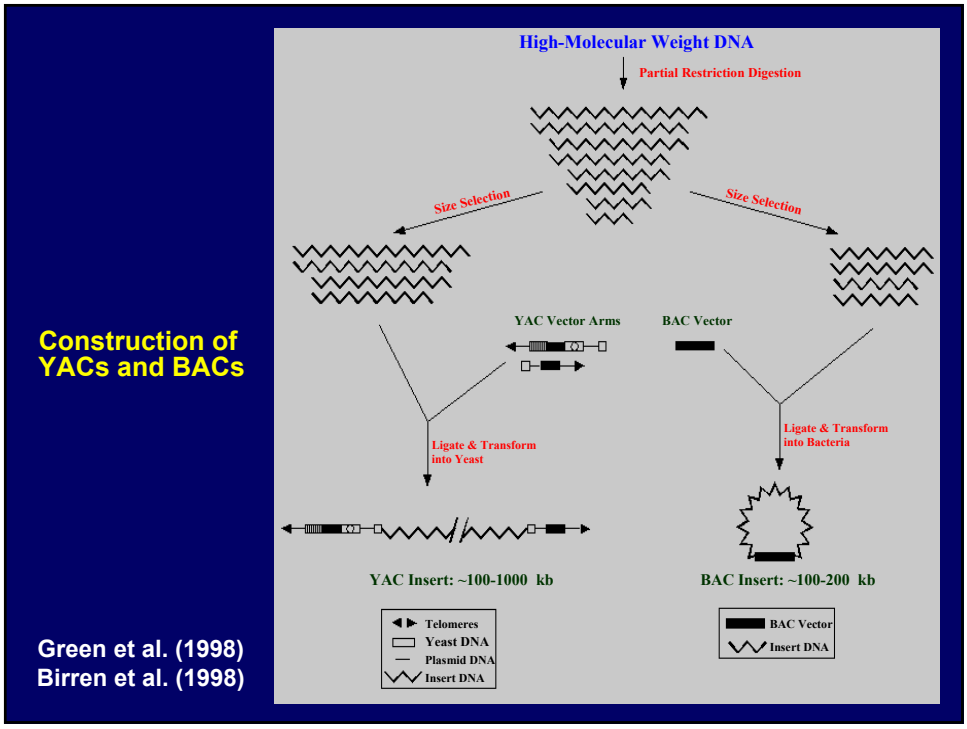
Sequence-Tagged Sites (STSs)

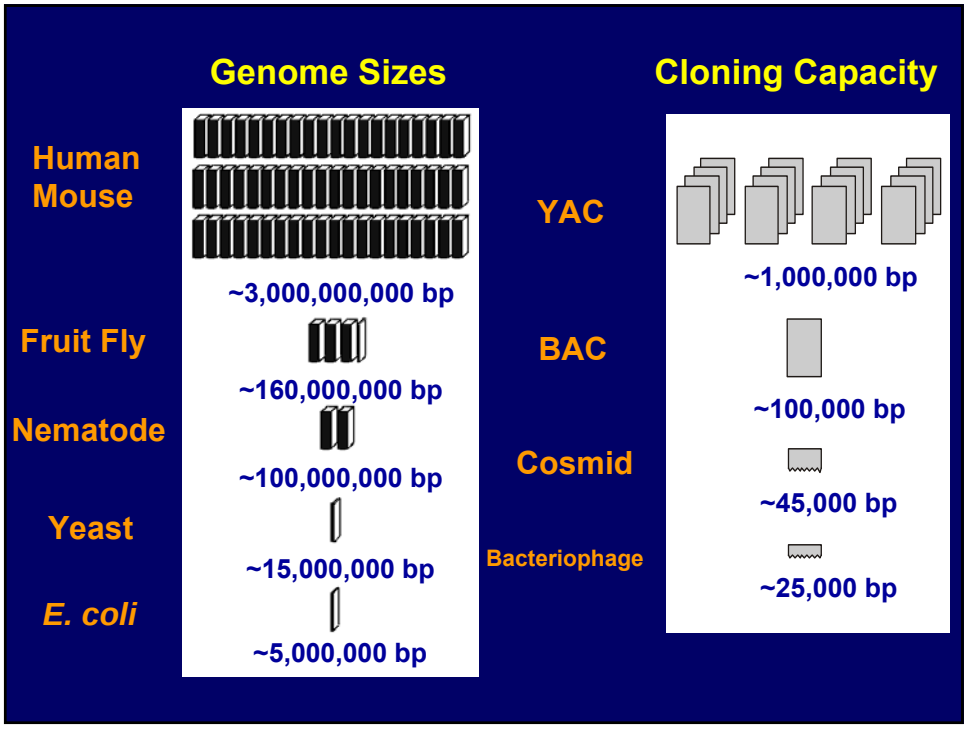


Physical Mapping: General Principles

- Importance of Physical Maps:
  - Localization and Isolation of Genes (e.g., Positional Cloning)
  - Study of Genome Organization and Evolution
  - Framework for Genome Sequencing
- Physical Mapping Involves Ordering Clones and/or Landmarks
- General Types of Physical Maps:
  - Landmark Only (e.g., Radiation Hybrid Maps)
  - Clone-Based
  - Sequence


Clone-Based Physical Mapping
Chromosome
Clones
Contigs

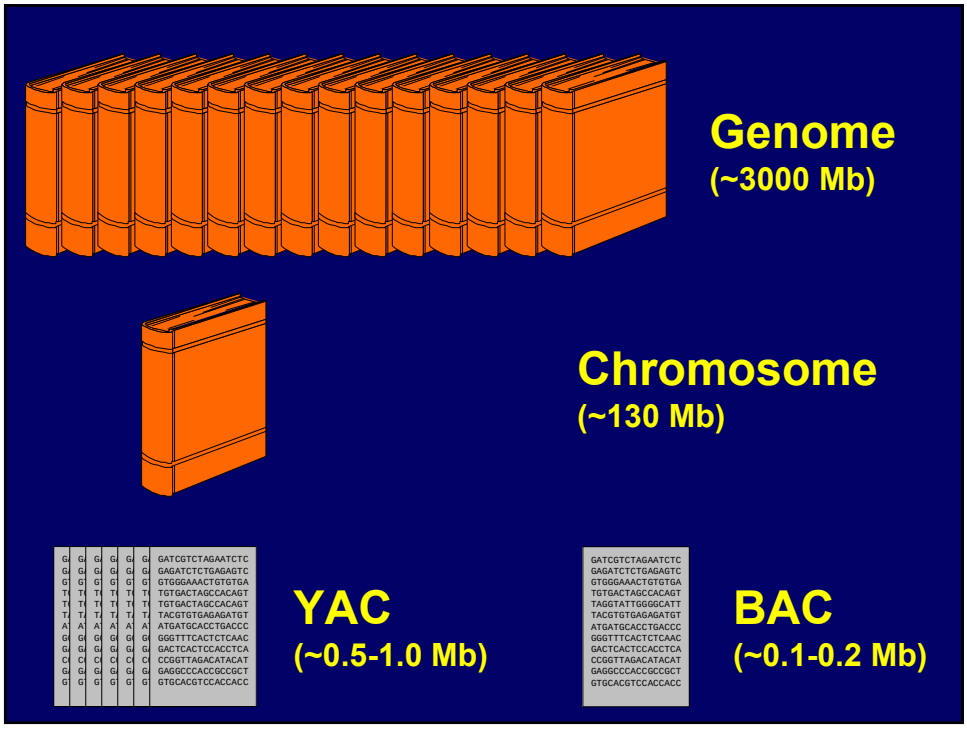

Construction of YACs and BACs
High-Molecular Weight DNA
Partial Restriction Digestion
Size Selection
Size Selection
YAC Vector Arms
BAC Vector
Ligate & Transform into Yeast
Ligate & Transform into Bacteria
YAC Insert: ~100-1000 kb
BAC Insert: ~100-200 kb
Telomeres
Yeast DNA
Plasmid DNA
Insert DNA
BAC Vector
Insert DNA
Green et al. (1998)
Birren et al. (1998)

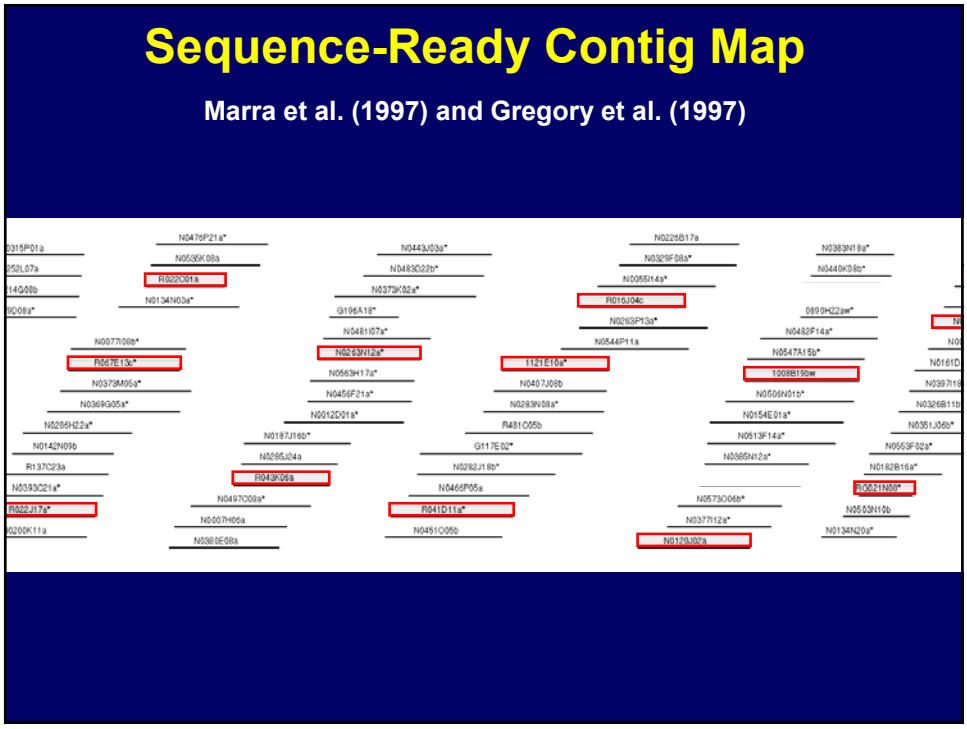


Bacterial Artificial Chromosomes (BACs)

- Bacterial-Based Cloning System Developed by Shizuya et al. (1992)
- Based on the *E. coli* F Factor (Fertility Plasmid): Replication Control
- Cloned Inserts: 100-200 kb, Circular DNA
- Low Copy Number

  Low Yields of DNA by Standard Methods
  Reasonably Stable
- Relatively Non-Chimeric
- BAC Libraries from Many Different Species now Available
  (e.g., www.chori.org/bacpac)
- See Birren et al. (1998)

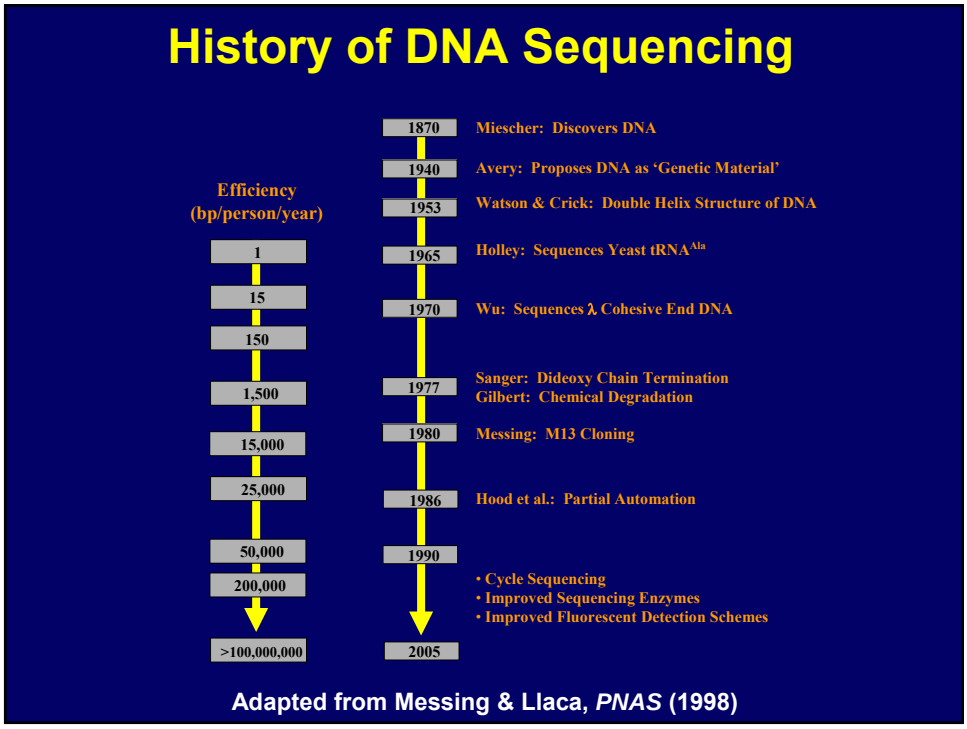

Genome
(~3000 Mb)
Chromosome
(~130 Mb)
YAC
(~0.5-1.0 Mb)
BAC
(~0.1-0.2 Mb)

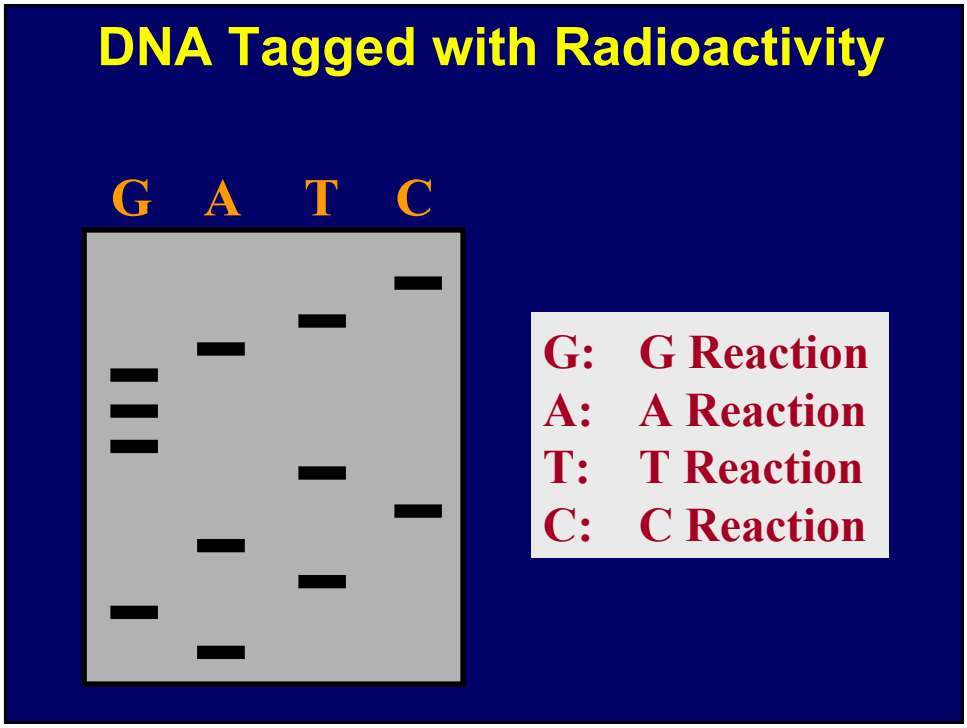

Sequence-Ready Contig Map
Marra et al. (1997) and Gregory et al. (1997)

Physical Mapping: Future Prospects

- Strategies for Physical Mapping are Radically Changing in the Sequence-Based Era
- Will Now See a Closer Interplay of Mapping and Sequencing in the Exploration of New Genomes
- Construction of New BAC Libraries will Allow Physical Mapping Studies of More Species' Genomes

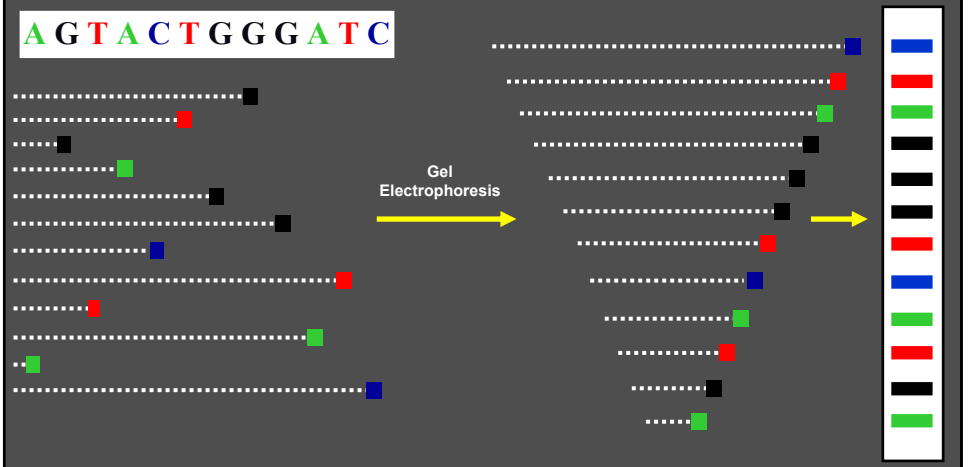
DNA Sequencing


History of DNA Sequencing
Efficiency
(bp/person/year)
1
15
150
1,500
15,000
25,000
50,000
200,000
>100,000,000
1870
1940
1953
1965
1970
1977
1980
1986
1990
2005
Miescher: Discovers DNA
Avery: Proposes DNA as 'Genetic Material'
Watson & Crick: Double Helix Structure of DNA
Holley: Sequences Yeast tRNA Ala
Wu: Sequences λ Cohesive End DNA
Sanger: Dideoxy Chain Termination
Gilbert: Chemical Degradation
Messing: M13 Cloning
Hood et al.: Partial Automation
• Cycle Sequencing
• Improved Sequencing Enzymes
• Improved Fluorescent Detection Schemes
Adapted from Messing & Llaca, PNAS (1998)

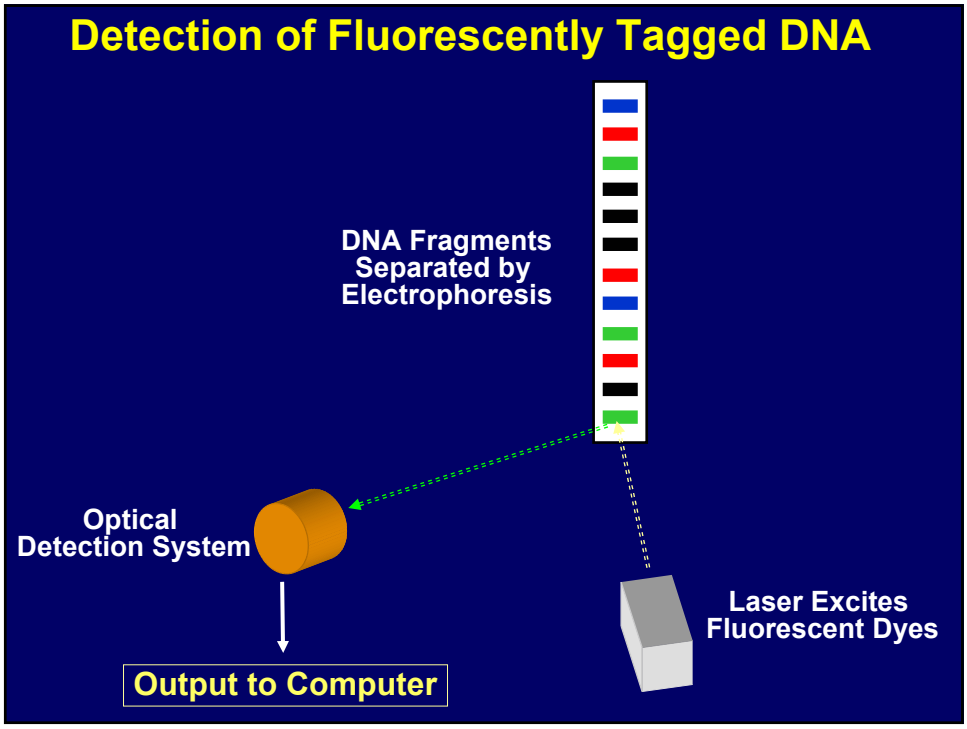

DNA Tagged with Radioactivity
G A T C
G: G Reaction
A: A Reaction
T: T Reaction
C: C Reaction

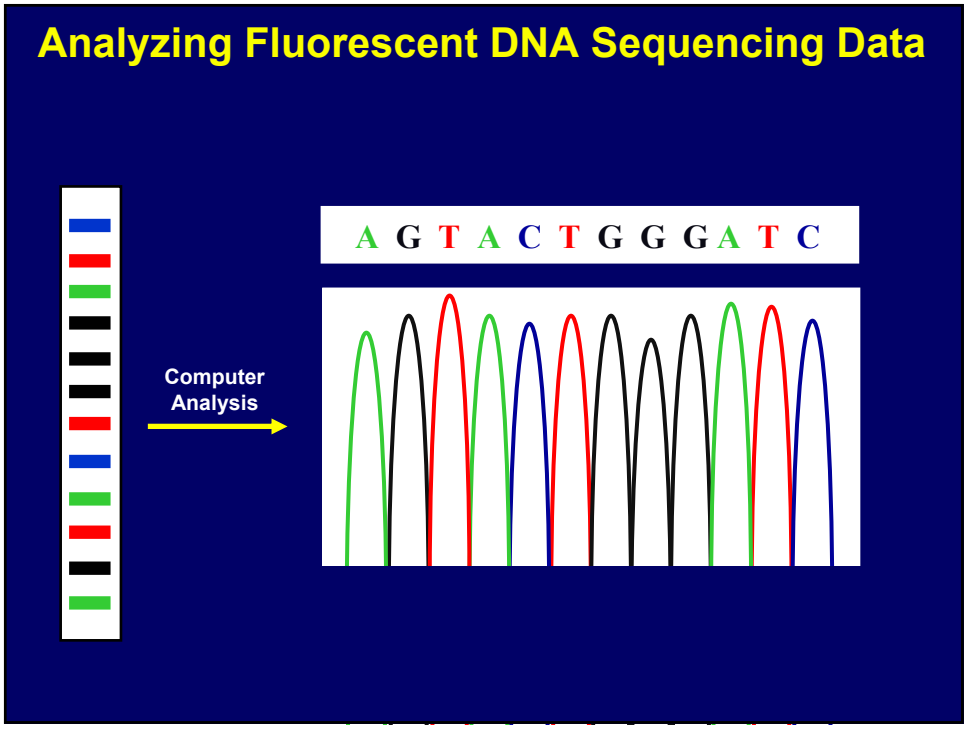
Radioactive Sequencing

Fluorescent DNA Sequencing

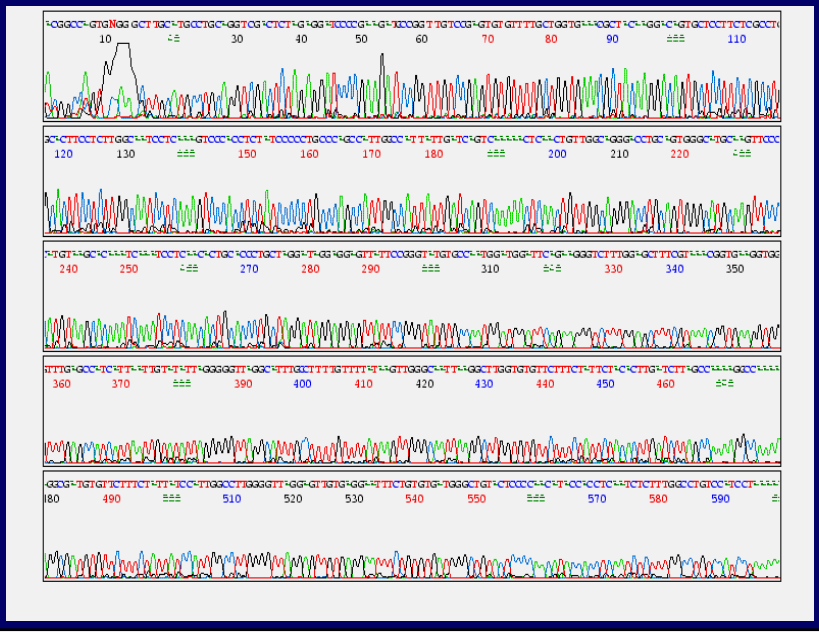


Wilson & Mardis (1997)


Detection of Fluorescently Tagged DNA
DNA Fragments Separated by Electrophoresis
Optical Detection System
Laser Excites Fluorescent Dyes
Output to Computer


Analyzing Fluorescent DNA Sequencing Data
Computer Analysis
A G T A C T G G G A T C

Fluorescent DNA Sequencing Results



Slab Gel-Based DNA Sequencing Instruments

Capillary-Based DNA Sequencing Instruments



Large-Scale cDNA Sequencing

- ESTs: Expressed-Sequence Tags
- SAGE: Serial Analysis of Gene Expression
- Full-Insert (Full-Length) cDNA Sequencing

Mammalian Gene Collection

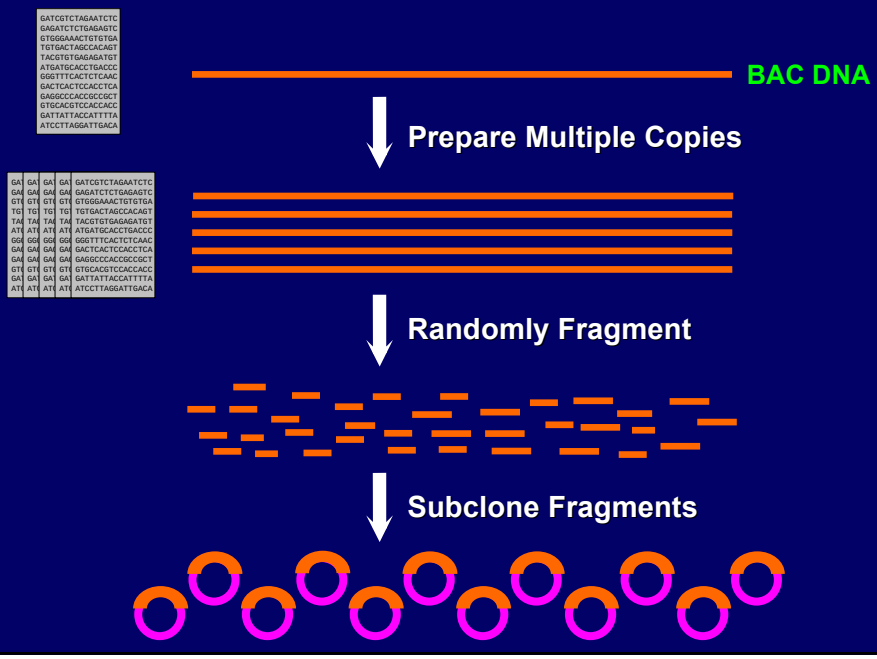
mgc.nci.nih.gov

Large-Scale Genomic Sequencing

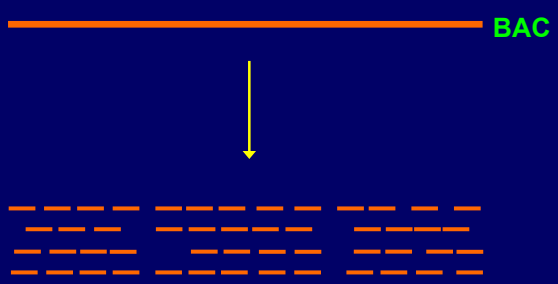
Shotgun Sequencing

Wilson & Mardis (1997)
Green (2001)

Subclone Construction



Shotgun Sequencing Strategy



Poisson Calculations

The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution

The probability a base is not sequenced is given by:

$$P_0 = e^{-c}$$

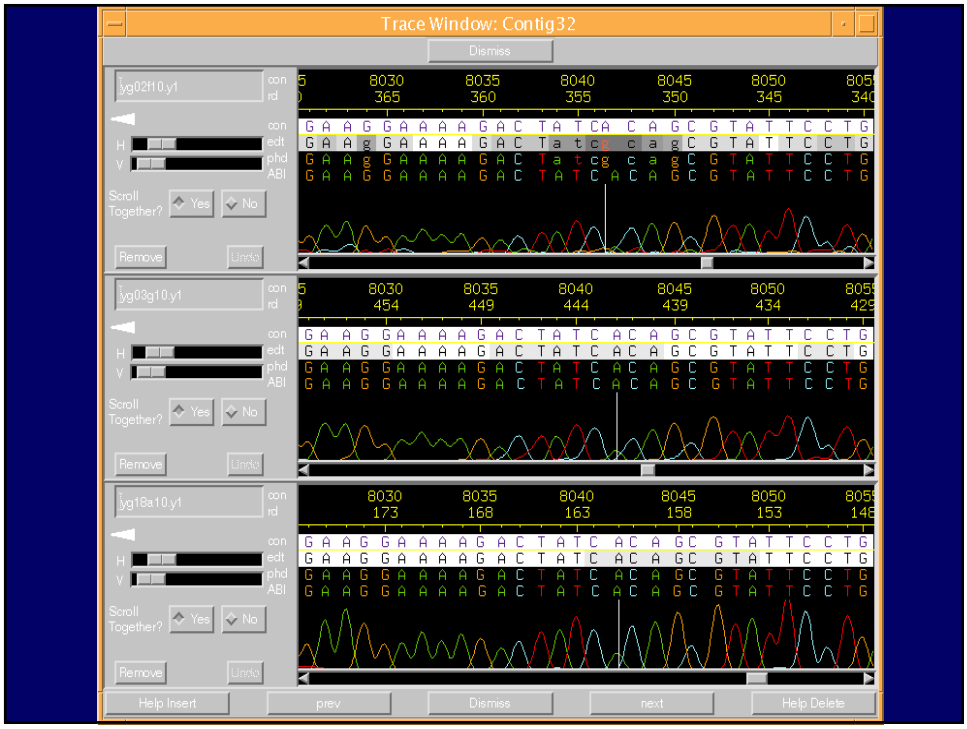
Where:

- c = fold sequence coverage (c=LN/G),
- LN = # bases sequenced, i.e. L = average sequencing read length and N = # reads
- G = target sequence length
- e = 2.718 (e=2.718281828459)

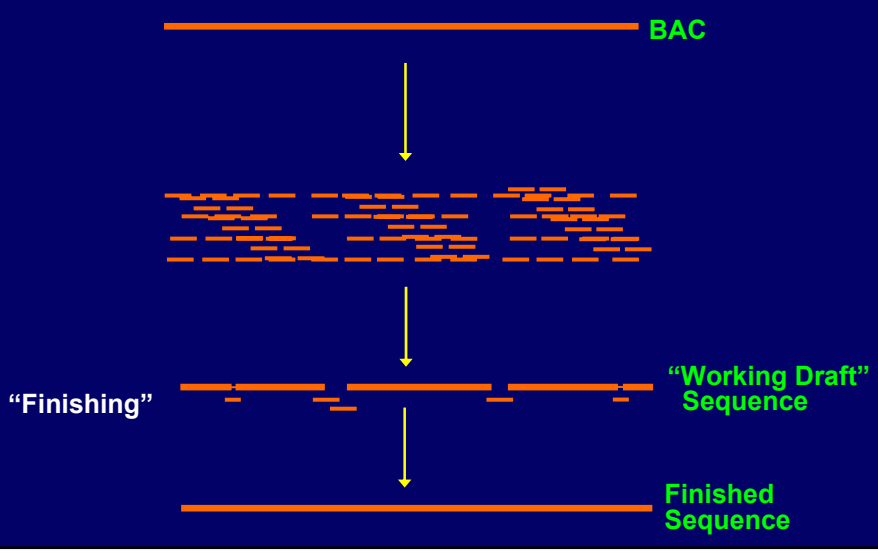
Fold Coverage	$P_0=e^{-c}$	% not sequenced	% sequenced
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

Shotgun Sequence Assembly

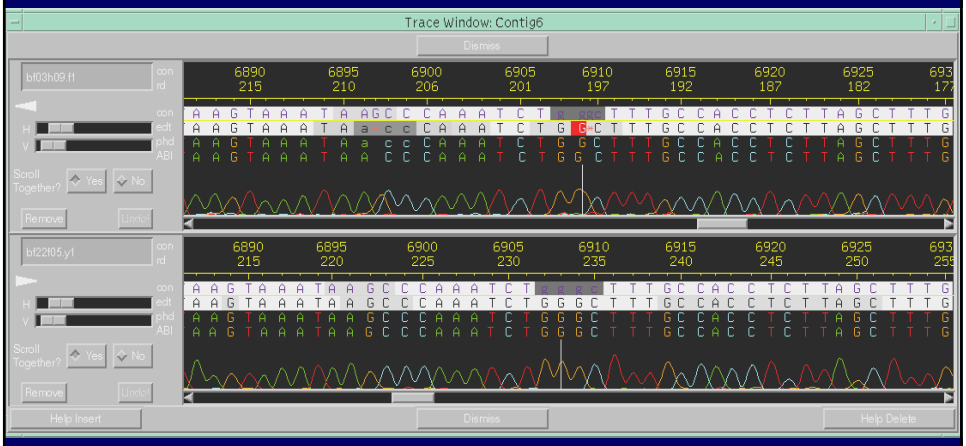
"Consed" (Gordon et al., 1998)



Shotgun Sequencing Strategy



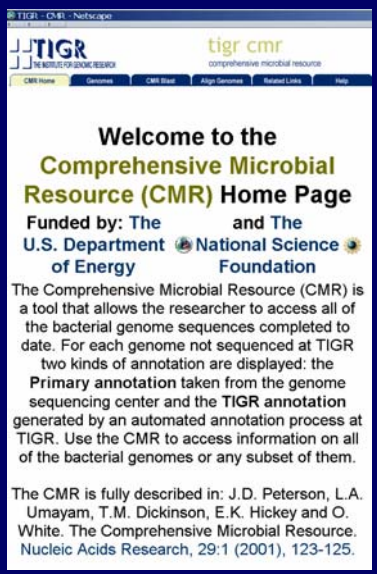
Sequence Finishing: Resolving Ambiguities



*** Sequence Finishing: Remains Relatively Expensive ***

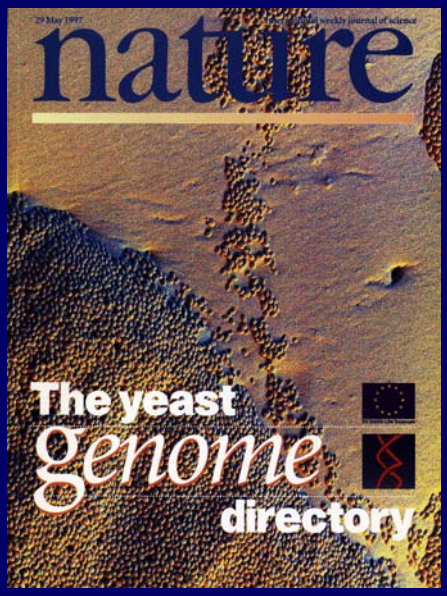
Historically Significant Genome Sequencing Projects

Bacterial Genome Sequences



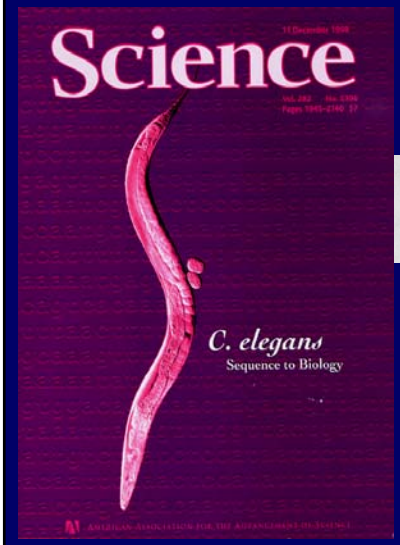
www.tigr.org

First Eukaryotic Genome Sequence



Nature 387:1-105, 1997

First Animal Genome Sequence

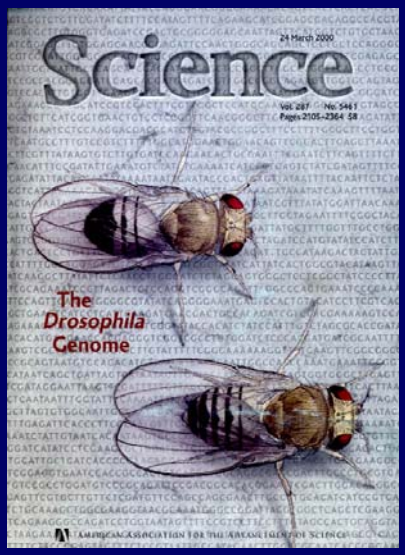


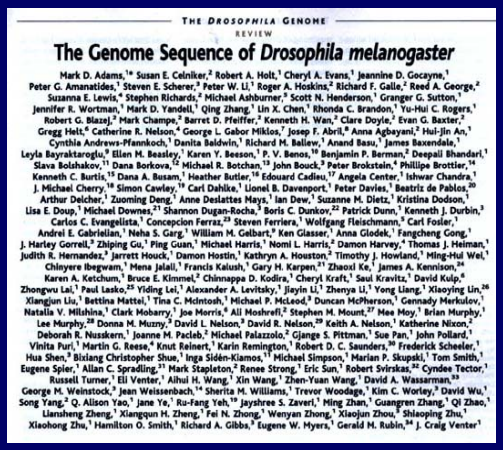
Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

Science 282:1012-2018, 1998

Second Animal Genome Sequence





THE DROSOPHILA GENOME

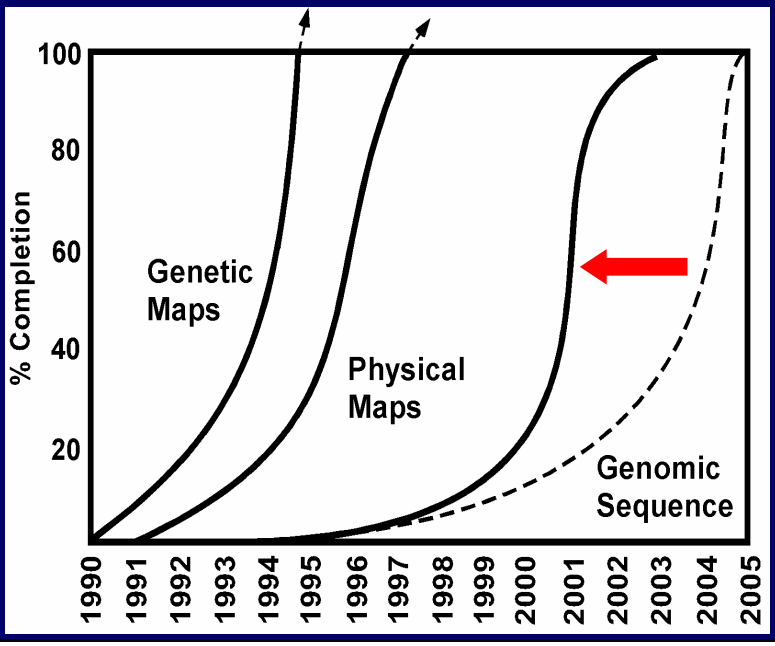
REVIEW

The Genome Sequence of *Drosophila melanogaster*

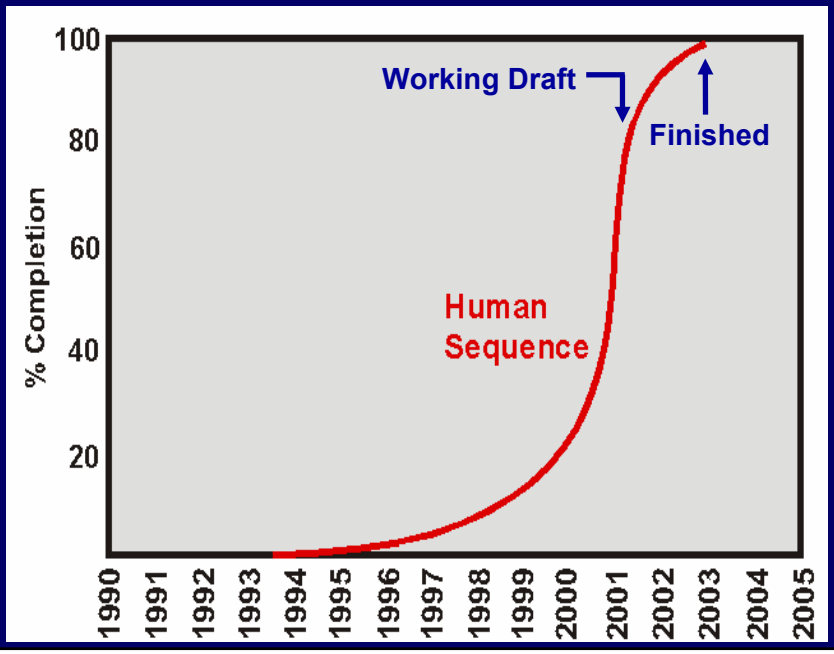
Mark D. Adams,[1*] Susan E. Celniker,[2] Robert A. Holt,[1] Cheryl A. Evans,[1] Jeannine D. Gocayne,[1] Peter G. Amanatides,[1] Steven E. Scherer,[1] Peter W. Li,[1] Roger A. Hoskins,[2] Richard F. Galle,[2] Reed A. George,[2] Suzanna E. Lewis,[2] Stephen Richards,[2] Michael Ashburner,[2] Scott N. Henderson,[1] Granger G. Sutton,[1] Jennifer R. Wortman,[1] Mark D. Yandell,[1] Qing Zhang,[1] Lin X. Chen,[1] Rhonda C. Brandon,[1] Yu-Hui C. Rogers,[1] Robert G. Blazej,[2] Mark Champe,[2] Barret D. Pfeiffer,[2] Kenneth H. Wan,[2] Clare Doyle,[2] Evan G. Baxter,[2] Gregg Helt,[2] Catherine R. Nelson,[2] George L. Gabor Miklos,[2] Josep F. Abril,[2] Anna Agbayani,[2] Hui-Jin An,[1] Cynthia Andrews-Pfannkoch,[1] Danita Baldwin,[1] Richard M. Ballew,[1] Anand Basu,[1] James Baxendale,[1] Leyla Bayraktaroglu,[2] Ellen M. Beasley,[1] Karen Y. Beeson,[1] P. V. Benos,[10] Benjamin P. Berman,[2] Deepali Bhandari,[1] Slava Bolshakov,[11] Dana Borkova,[12] Michael R. Botchan,[13] John Bouck,[2] Peter Brokstein,[2] Phillipe Brottier,[14] Kenneth C. Burtis,[15] Dana A. Busam,[1] Heather Butler,[16] Edouard Cadieu,[17] Angela Center,[1] Ishwar Chandra,[1] J. Michael Cherry,[18] Simon Cawley,[19] Carl Dahlke,[1] Lionel B. Davenport,[1] Peter Davies,[1] Beatriz de Pablos,[20] Arthur Delcher,[1] Zuoming Deng,[1] Anne Deslattes Mays,[1] Ian Dew,[1] Suzanne M. Dietz,[1] Kristina Dodson,[1] Lisa E. Doup,[1] Michael Downes,[21] Shannon Dugan-Rocha,[1] Boris C. Dunkov,[22] Patrick Dunn,[1] Kenneth J. Durbin,[1] Carlos C. Evangelista,[1] Concepcion Ferraz,[23] Steven Ferriera,[1] Wolfgang Fleischmann,[1] Carl Fosler,[1] Andrei E. Gabrielian,[1] Neha S. Garg,[1] William M. Gelbart,[9] Ken Glasser,[1] Anna Glodek,[1] Fangcheng Gong,[1] J. Harley Gorrell,[2] Zhiping Gu,[1] Ping Guan,[1] Michael Harris,[1] Nomi L. Harris,[2] Damon Harvey,[4] Thomas J. Heiman,[1] Judith R. Hernandez,[1] Jarrett Houck,[1] Damon Hostin,[1] Kathryn A. Houston,[2] Timothy J. Howland,[1] Ming-Hui Wei,[1] Chinyere Ibegwam,[1] Mena Jalali,[1] Francis Kalush,[1] Gary H. Karpen,[21] Zhaoxi Ke,[1] James A. Kennison,[24] Karen A. Ketchum,[1] Bruce E. Kimmel,[2] Chinnappa D. Kodira,[1] Cheryl Kraft,[1] Saul Kravitz,[1] David Kulp,[4] Zhongwu Lai,[1] Paul Lasko,[25] Yiding Lei,[1] Alexander A. Levitsky,[1] Jiayin Li,[1] Zhenya Li,[1] Yong Liang,[1] Xiaoying Lin,[26] Xiangjun Liu,[1] Bettina Mattei,[1] Tina C. McIntosh,[1] Michael P. McLeod,[3] Duncan McPherson,[1] Gennady Merkulov,[1] Natalia V. Milshina,[1] Clark Mobarry,[1] Joe Morris,[4] Ali Moshrefi,[2] Stephen M. Mount,[27] Mee Moy,[1] Brian Murphy,[1] Lee Murphy,[28] Donna M. Muzny,[3] David L. Nelson,[3] David R. Nelson,[29] Keith A. Nelson,[1] Katherine Nixon,[2] Deborah R. Nusskern,[1] Joanne M. Pacleb,[2] Michael Palazzolo,[2] Gjange S. Pittman,[1] Sue Pan,[1] John Pollard,[1] Vinita Puri,[1] Martin G. Reese,[4] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[3] Bixiang Christopher Shue,[1] Inga Sidén-Kiamos,[31] Michael Simpson,[1] Marian P. Skupski,[1] Tom Smith,[1] Eugene Spier,[1] Allan C. Spradling,[31] Mark Stapleton,[2] Renee Strong,[1] Eric Sun,[1] Robert Svirskas,[32] Cyndee Tector,[1] Russell Turner,[1] Eli Venter,[1] Alhul H. Wang,[1] Xin Wang,[1] Zhen-Yuan Wang,[1] David A. Wassarman,[33] George M. Weinstock,[3] Jean Weissenbach,[34] Sherita M. Williams,[1] Trevor Woodage,[1] Kim C. Worley,[3] David Wu,[1] Song Yang,[2] Q. Alison Yao,[1] Jane Ye,[1] Ru-Fang Yeh,[10] Jayshree S. Zaveri,[1] Ming Zhan,[1] Guangren Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Xiangqun H. Zheng,[1] Fei N. Zhong,[1] Wenyan Zhong,[1] Xiaojun Zhou,[2] Shiaoping Zhu,[1] Xiaohong Zhu,[1] Hamilton O. Smith,[1] Richard A. Gibbs,[3] Eugene W. Myers,[1] Gerald M. Rubin,[24] J. Craig Venter[1]

Science 287:2185-2195, 2000

Revised Timetable for Human Genome Sequencing

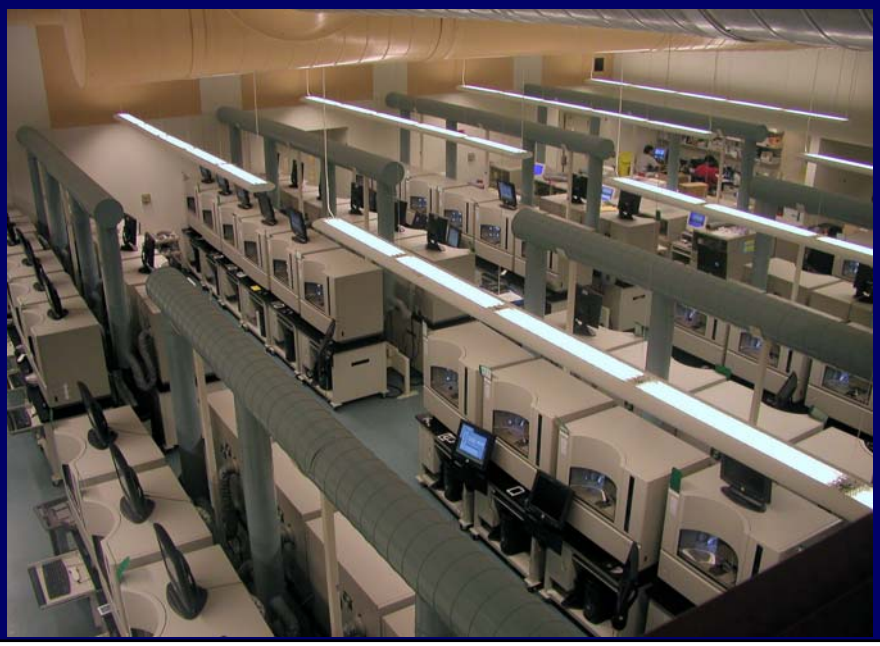


Timetable for Human Genome Sequencing



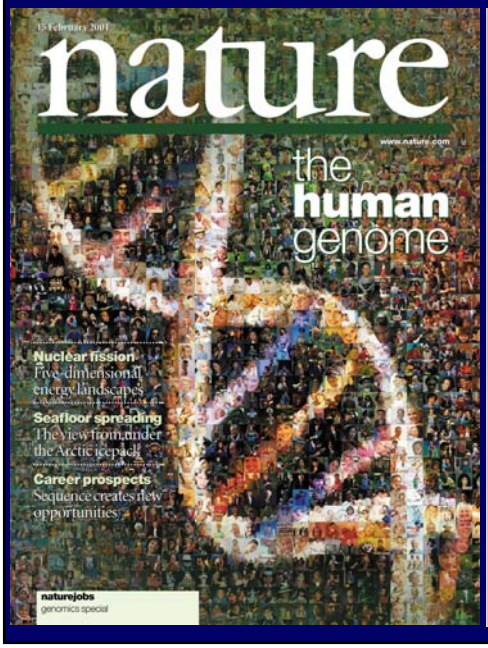
Human Genome Sequencing Centers

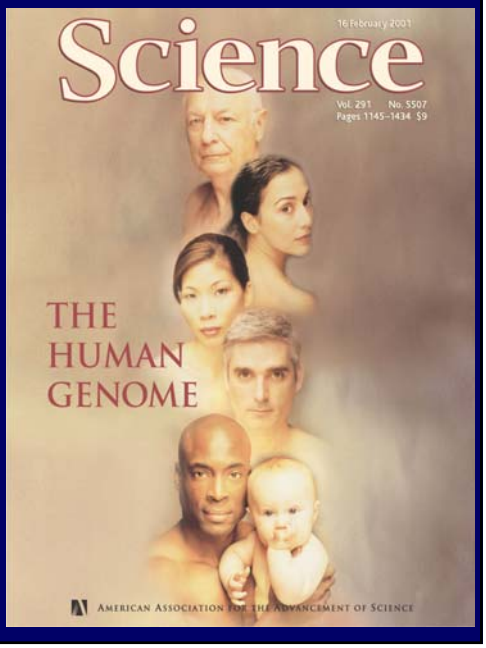
Human Genome Sequencing Centers



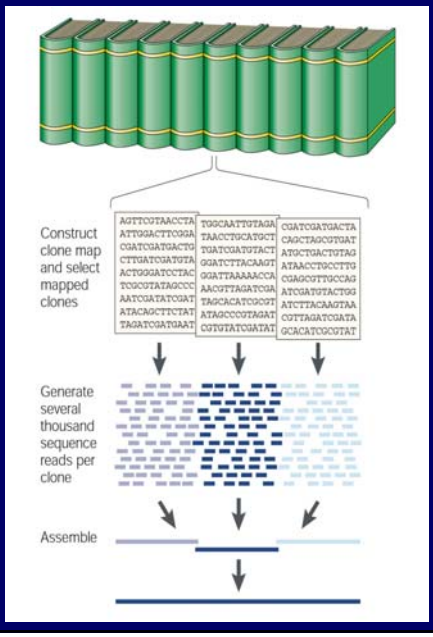
June, 2000 Announcement

February, 2001 Publications



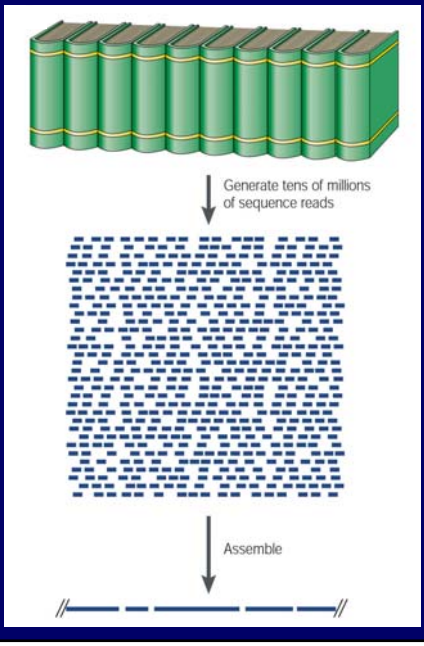


BAC-by-BAC Shotgun Sequencing



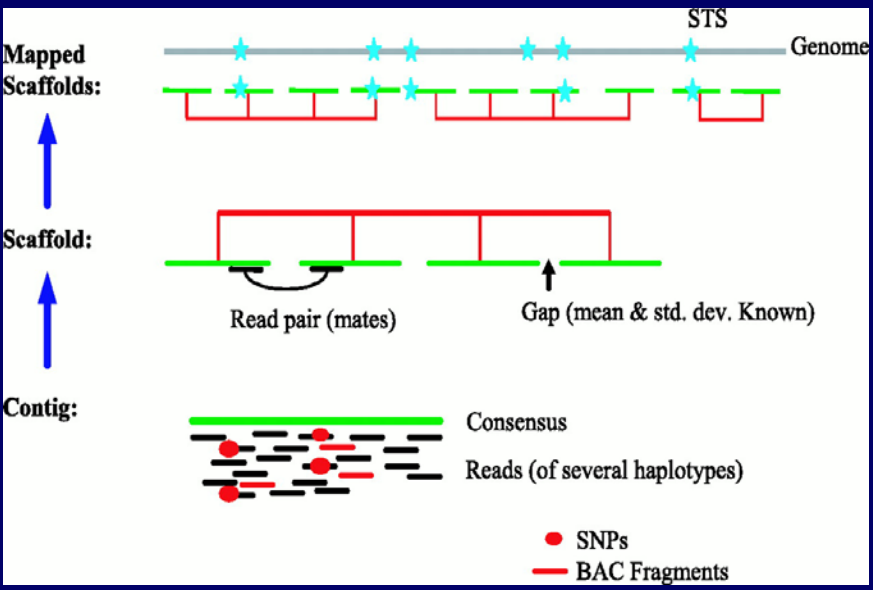
Green (2001)

Whole-Genome Shotgun Sequencing



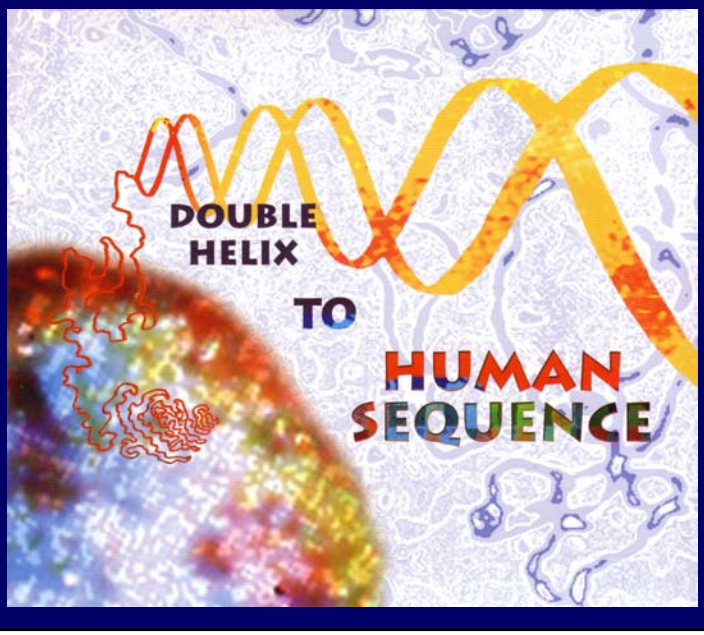
Green (2001)

Whole-Genome Shotgun Sequence Assembly



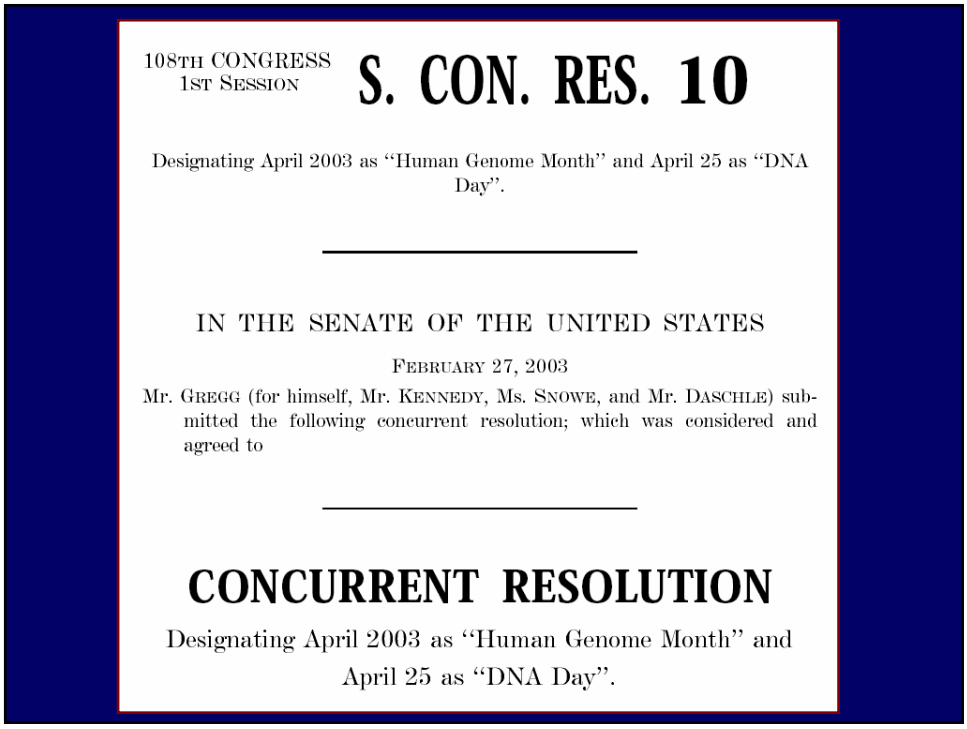
Venter et al., 2001

April, 2003 Completion



International Human Genome Sequencing Consortium

- 6 Countries
- 20 Sequencing Centers
- 1000's of Individuals
- ~1,000 bases per second, 24 hours per day, 7 days per week



108TH CONGRESS
1ST SESSION

S. CON. RES. 10

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".

IN THE SENATE OF THE UNITED STATES

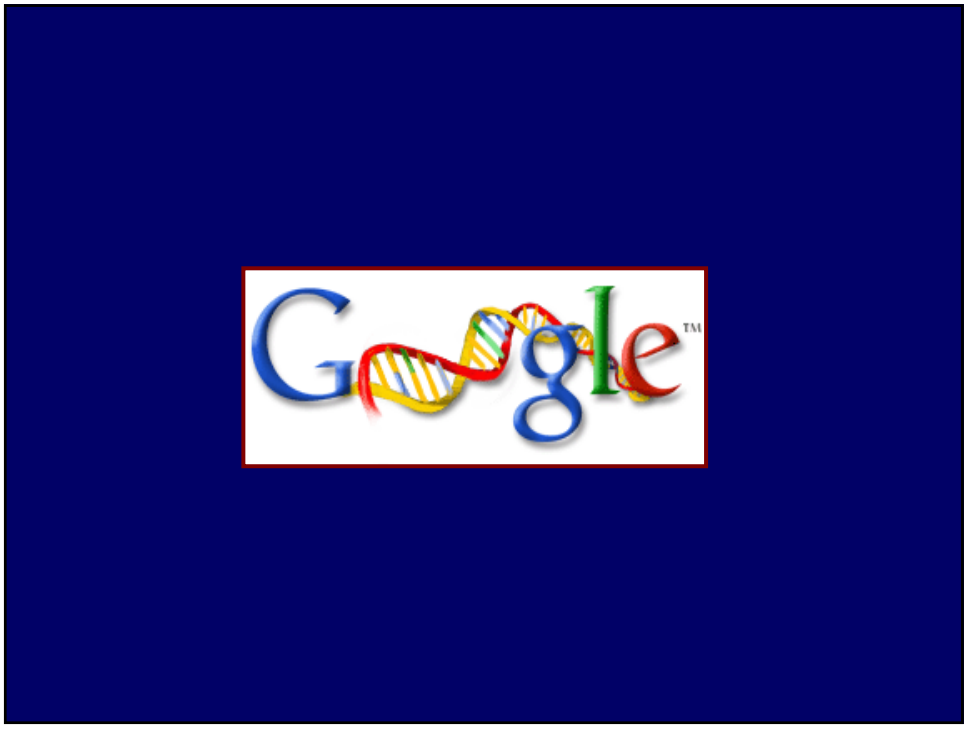
FEBRUARY 27, 2003

Mr. GREGG (for himself, Mr. KENNEDY, Ms. SNOWE, and Mr. DASCHLE) submitted the following concurrent resolution; which was considered and agreed to

---

CONCURRENT RESOLUTION

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".


Google™


2nd
Genome The End of the Beginning

October, 2004 Publication


articles

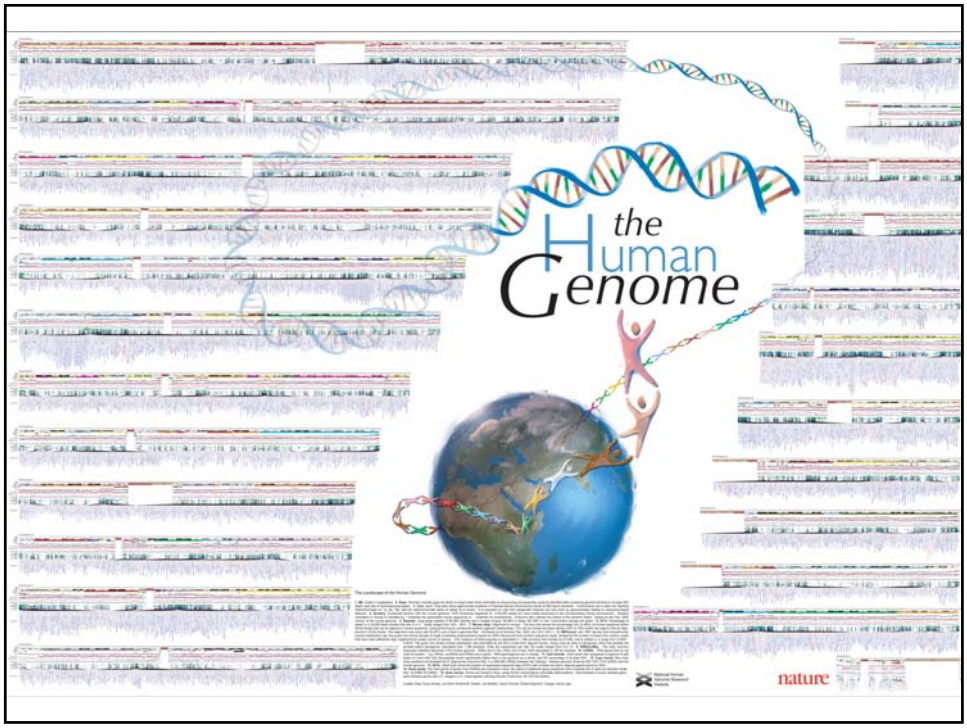
Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

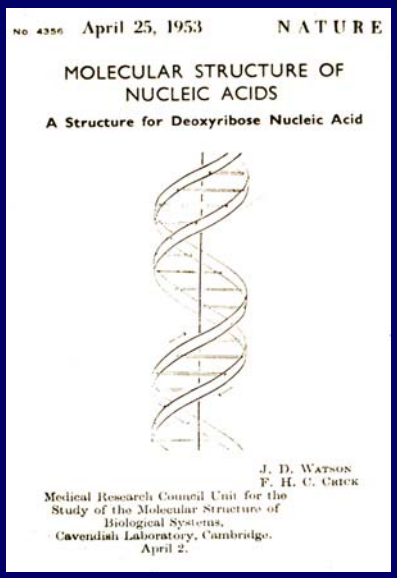
*A list of authors and their affiliations appears in the Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of bioinformatic analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

Nature 431:931-945, 2004



April, 1953 → **April, 2003**



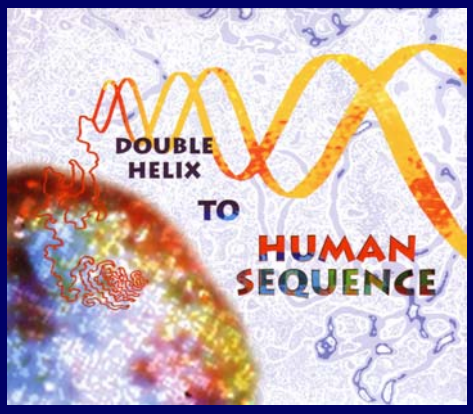
No 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

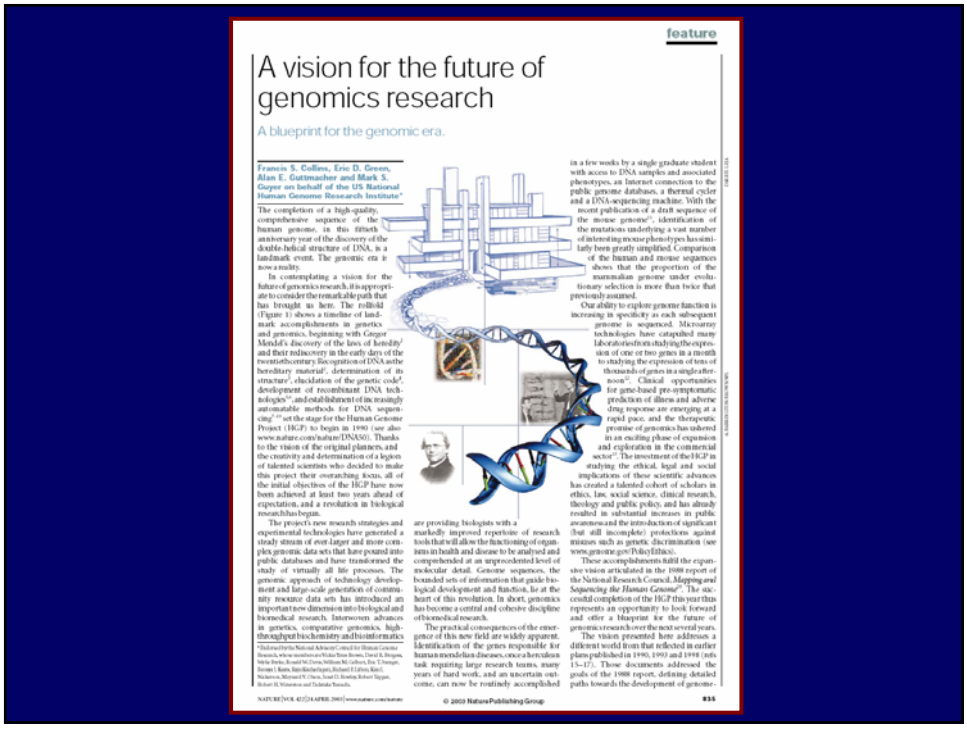
J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.



All of the original goals of the Human Genome Project have been accomplished!

What's Next?



feature

A vision for the future of genomics research

A blueprint for the genomic era.

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute*

The completion of a high-quality, comprehensive sequence of the human genome, in this fiftieth anniversary year of the discovery of the double-helical structure of DNA, is a landmark event. The genomic era is now a reality.

In contemplating a vision for the future of genomics research, it is appropriate to consider the remarkable path that has brought us here. The rollfold (Figure 1) shows a timeline of landmark accomplishments in genetics and genomics, beginning with Gregor Mendel's discovery of the laws of heredity and their rediscovery in the early days of the twentieth century. Recognition of DNA as the hereditary material[1], determination of its structure[2], elucidation of the genetic code[3], development of recombinant DNA technologies[4–6], and establishment of increasingly automatable methods for DNA sequencing[7–10] set the stage for the Human Genome Project (HGP) to begin in 1990 (see also www.nature.com/nature/DNA50). Thanks to the vision of the original planners, and the creativity and determination of a legion of talented scientists who decided to make this project their overarching focus, all of the initial objectives of the HGP have now been achieved at least two years ahead of expectation, and a revolution in biological research has begun.

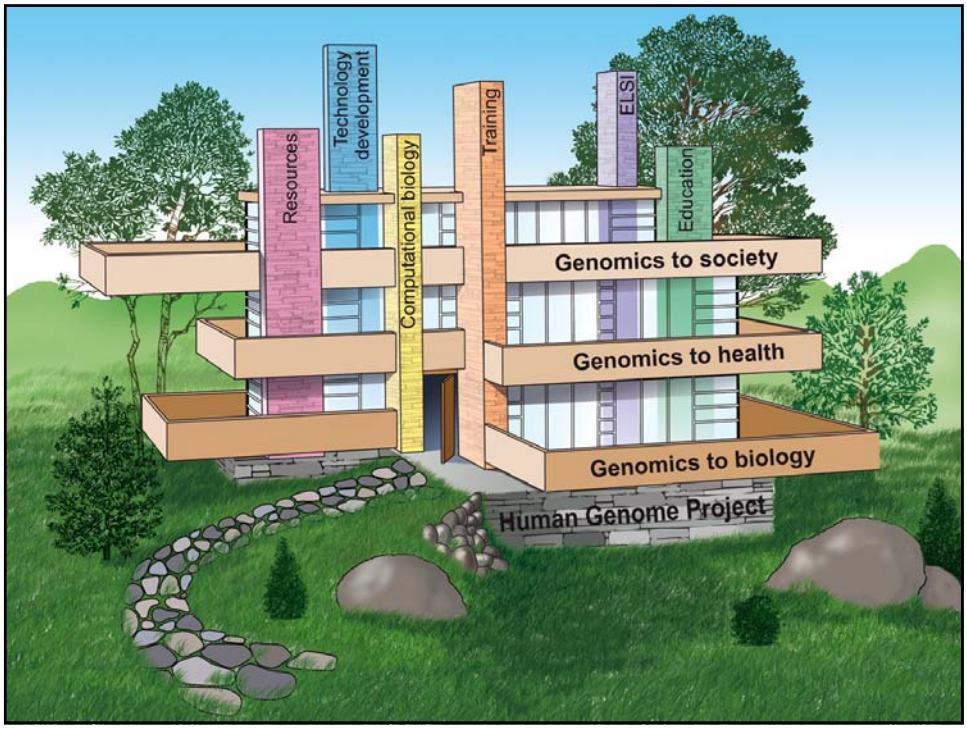
The project's new research strategies and experimental technologies have generated a steady stream of ever-larger and more complex genomic data sets that have poured into public databases and have transformed the study of virtually all life processes. The genomic approach of technology development and large-scale generation of community resource data sets has introduced an important new dimension into biological and biomedical research. Interwoven advances in genetics, comparative genomics, high-throughput biochemistry and bioinformatics are providing biologists with a markedly improved repertoire of research tools that will allow the functioning of organisms in health and disease to be analysed and comprehended at an unprecedented level of molecular detail. Genome sequences, the bounded sets of information that guide biological development and function, lie at the heart of this revolution. In short, genomics has become a central and cohesive discipline of biomedical research.

The practical consequences of the emergence of this new field are widely apparent. Identification of the genes responsible for human mendelian diseases, once a herculean task requiring large research teams, many years of hard work, and an uncertain outcome, can now be routinely accomplished in a few weeks by a single graduate student with access to DNA samples and associated phenotypes, an Internet connection to the public genome databases, a thermal cycler and a DNA-sequencing machine. With the recent publication of a draft sequence of the mouse genome[11], identification of the mutations underlying a vast number of interesting mouse phenotypes has similarly been greatly simplified. Comparison of the human and mouse sequences shows that the proportion of the mammalian genome under evolutionary selection is more than twice that previously assumed.

Our ability to explore genome function is increasing in specificity as each subsequent genome is sequenced. Microarray technologies have catapulted many laboratories from studying the expression of one or two genes in a month to studying the expression of tens of thousands of genes in a single afternoon[12]. Clinical opportunities for gene-based pre-symptomatic prediction of illness and adverse drug response are emerging at a rapid pace, and the therapeutic promise of genomics has ushered in an exciting phase of expansion and exploration in the commercial sector[13]. The investment of the HGP in studying the ethical, legal and social implications of these scientific advances has created a talented cohort of scholars in ethics, law, social science, clinical research, theology and public policy, and has already resulted in substantial increases in public awareness and the introduction of significant (but still incomplete) protections against misuses such as genetic discrimination (see www.genome.gov/PolicyEthics).

These accomplishments fulfill the expansive vision articulated in the 1988 report of the National Research Council, *Mapping and Sequencing the Human Genome*[14]. The successful completion of the HGP this year thus represents an opportunity to look forward and offer a blueprint for the future of genomics research over the next several years.

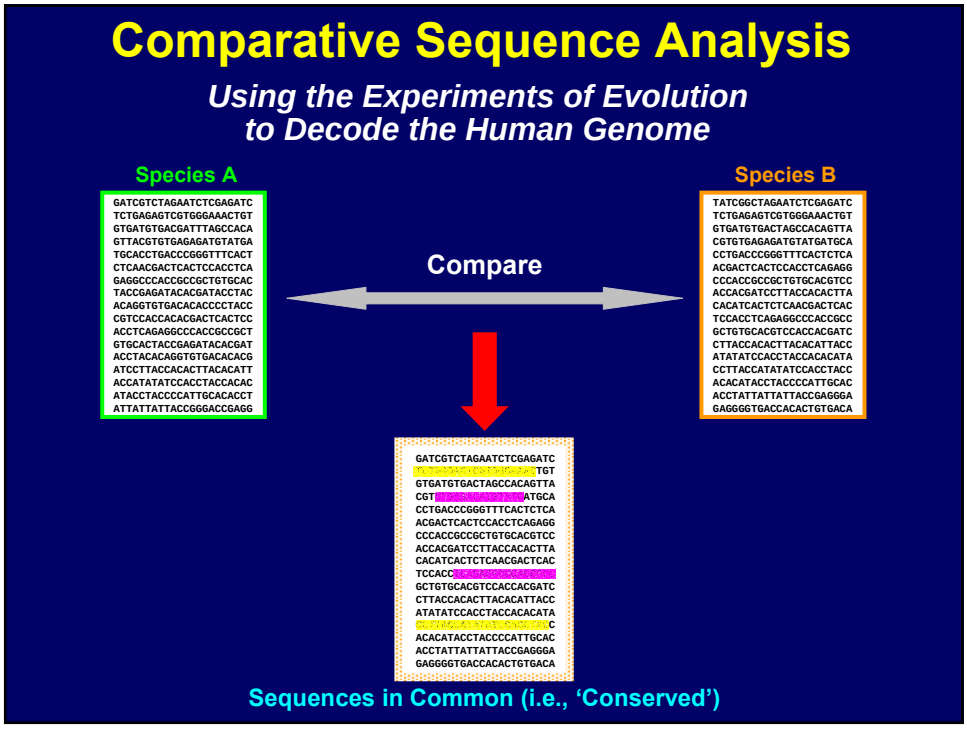
The vision presented here addresses a different world from that reflected in earlier plans published in 1990, 1993 and 1998 (refs 15–17). Those documents addressed the goals of the 1988 report, defining detailed paths towards the development of genome-

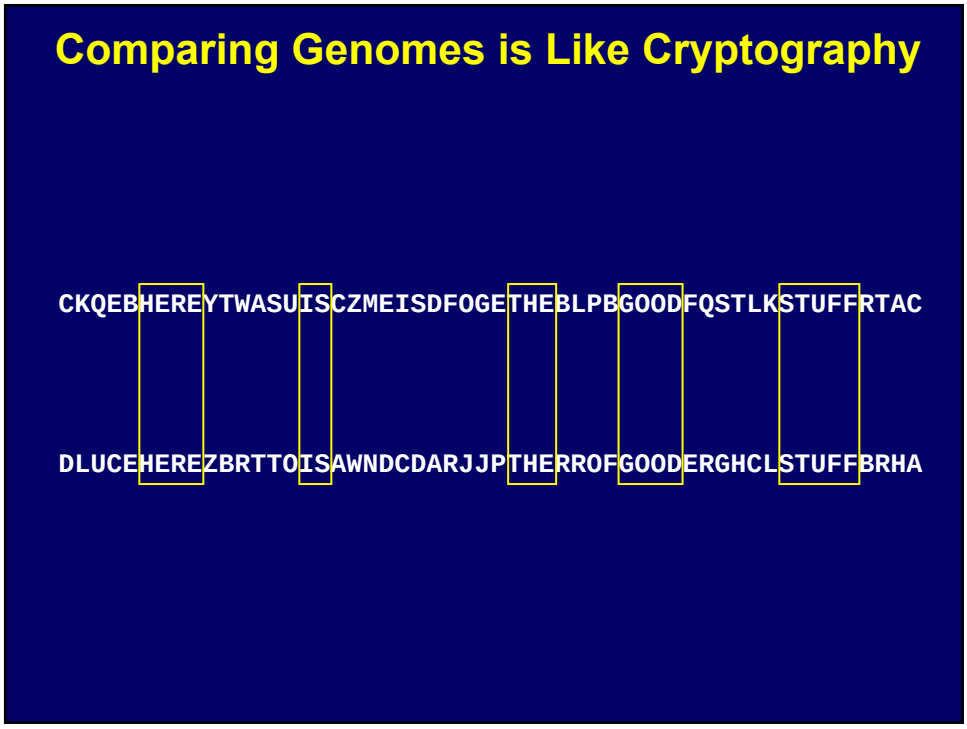


Mapping the Human Genome	~1990 to ~2000	The Human Genome Project
Sequencing the Human Genome	~1998 to ~2003	
Interpreting the Human Genome Sequence	~2003 to ???	*Beyond* The Human Genome Project

~3,000 bp (0.0001%) of Human Genome Sequence

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCCACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTTGGGGCCGGACCAGGCAGCACTCGGCT[redacted]AGAGCAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGGTGGGTGGGGGG[redacted]
[redacted]GTGAGAAGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTTAAAAAGATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTTGTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGG
ACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATT
TTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTATTTGTCACAGGTAAATGATTCTTCAAAAATTG
AAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGGAATAGATACCGAAGTTATATCCAA
CTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAATAGCATCTAAGTTCGG
AATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTAC[redacted]AAAAAGTCAAATATGT
TATAAATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCC[redacted]TATCTTTTTGTGT
TGATGTTATCCACCTTTTGTTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCAT
AAATTATAGACTTTTTTTAGCAGAGAACTTTGAGGAACCTAAATGCCAAC[redacted]TTCATGGATA
GTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTGATCTGTCAAAGTGGGTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTT
AATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGTG[redacted]
[redacted]CCAATGA
TTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATTAAGTTAGGCAGTATTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTTAACATCTCTGTG
TGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATG
TTTCTATTGTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT


Comparative Sequence Analysis
Using the Experiments of Evolution
to Decode the Human Genome
Species A
GATCGTCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACGATTTAGCCACA
GTTACGTGTGAGAGATGTATGA
TGCACCTGACCCGGGTTTCACT
CTCAACGACTCACTCCACCTCA
GAGGCCCACCGCCGCTGTGCAC
TACCGAGATACACGATACCTAC
ACAGGTGTGACACACCCCTACC
CGTCCACCACACGACTCACTCC
ACCTCAGAGGCCCACCGCCGCT
GTGCACTACCGAGATACACGAT
ACCTACACAGGTGTGACACACG
ATCCTTACCACACTTACACATT
ACCATATATCCACCTACCACAC
ATACCTACCCCATTGCACACCT
ATTATTATTACCGGGACCGAGG
Compare
Species B
TATCGGCTAGAATCTCGAGATC
TCTGAGAGTCGTGGGAAACTGT
GTGATGTGACTAGCCACAGTTA
CGTGTGAGAGATGTATGATGCA
CCTGACCCGGGTTTCACTCTCA
ACGACTCACTCCACCTCAGAGG
CCCACCGCCGCTGTGCACGTCC
ACCACGATCCTTACCACACTTA
CACATCACTCTCAACGACTCAC
TCCACCTCAGAGGCCCACCGCC
GCTGTGCACGTCCACCACGATC
CTTACCACACTTACACATTACC
ATATATCCACCTACCACACATA
CCTTACCATATATCCACCTACC
ACACATACCTACCCCATTGCAC
ACCTATTATTATTACCGAGGGA
GAGGGGTGACCACACTGTGACA
GATCGTCTAGAATCTCGAGATC
TGT
GTGATGTGACTAGCCACAGTTA
ATGCA
CCTGACCCGGGTTTCACTCTCA
ACGACTCACTCCACCTCAGAGG
CCCACCGCCGCTGTGCACGTCC
ACCACGATCCTTACCACACTTA
CACATCACTCTCAACGACTCAC
GCTGTGCACGTCCACCACGATC
CTTACCACACTTACACATTACC
ATATATCCACCTACCACACATA
C
ACACATACCTACCCCATTGCAC
ACCTATTATTATTACCGAGGGA
GAGGGGTGACCACACTGTGACA
Sequences in Common (i.e., 'Conserved')

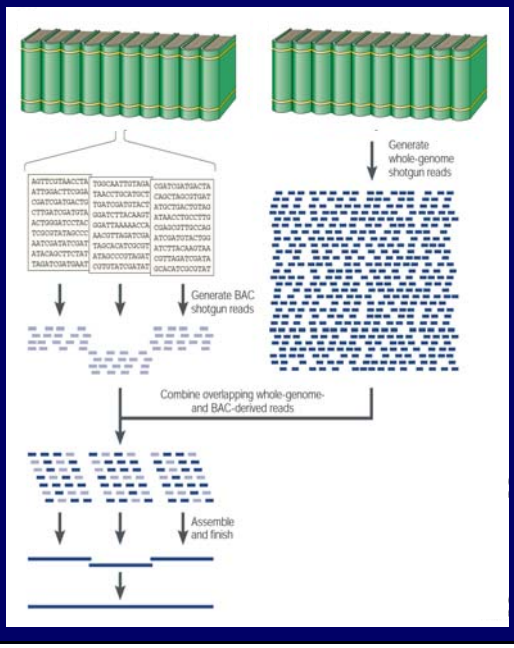

Comparing Genomes is Like Cryptography
CKQEBHEREYTWASUISCZMEISDFOGETHEBLPBGOODFQSTLKSTUFFRTAC
DLUCEHEREZBRTTOISAWNDCDARJJPTHERROFGOODERGHCLSTUFFBRHA

Functional Elements: Coding vs. Non-Coding

- **Coding Sequences (i.e., Genes)**
  - Relatively EASY to Identify
  - Mostly Know What to Look For
  - Complementary Data Sets Available (ESTs, cDNAs)
  - Ever-Improving Computational Gene Predictions
- **Non-Coding Functional Sequences**
  - HARD to Identify
  - Very Little Known About What to Look For
  - Virtually No Complementary Data Sets Available
  - Poor Computational Predictions

Major role for comparative sequence analysis will be the identification of functionally important, non-coding sequences

Hybrid Shotgun Sequencing



Green (2001)

Hybrid Shotgun Sequencing

What is the optimal mixture???

Redundant Coverage:	0	~5-fold	~10-fold
	Whole-Genome Shotgun Sequence Reads		
	BAC-by-BAC Shotgun Sequence Reads		
Redundant Coverage:	~10-fold	~5-fold	0

Nature 420:520-562, 2002

articles

Initial sequencing and comparative analysis of the mouse genome

Nature 428:493-521, 2004

articles

Genome sequence of the Brown Norway rat yields insights into mammalian evolution

Human-Rodent Sequence Comparisons

Nature
420:520-562, 2002

Nature
428:493-521, 2004

- ~40% in Alignments
- ~5% Under Selection
  - ~1.5% Protein Coding
  - ~3.5% Non-Coding

Multi-Species Sequence Comparisons

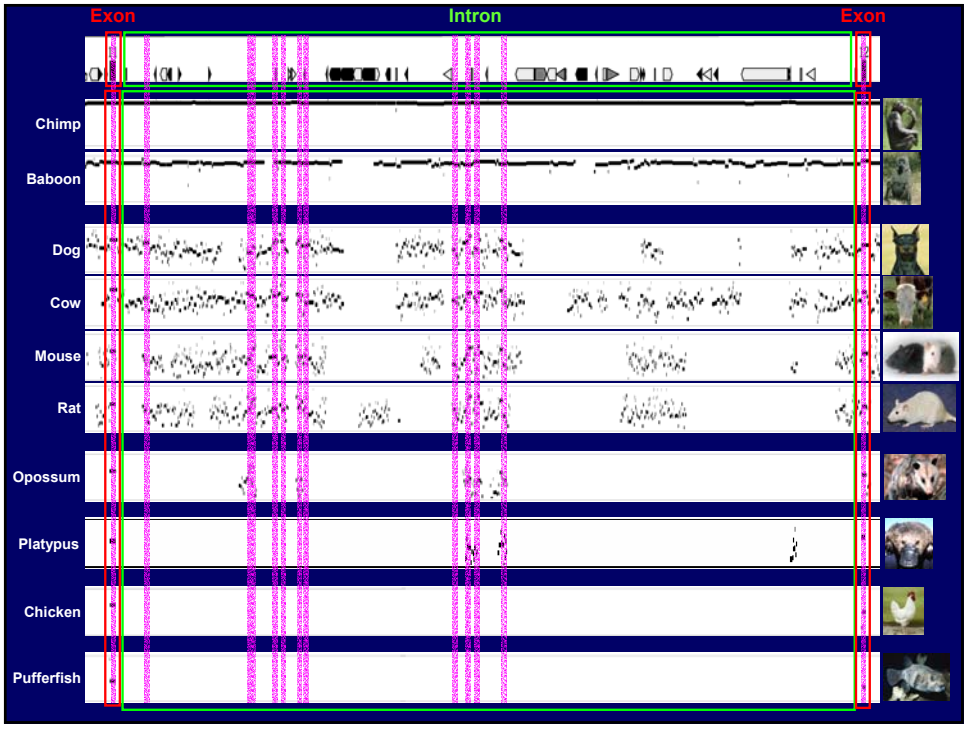
HUMAN

Multi-Species Comparative Sequence Analysis

Comparative analyses of multi-species sequences from targeted genomic regions

- Targeted Genomic Regions
- BAC-Based Sequencing in Multiple Vertebrates
- Identify Highly Conserved Non-Coding Sequences
- Conserved Sequences Correlate with Functional Elements

Thomas et al. (2003)



Additional Vertebrate Genome Sequencing Efforts


Chimpanzee


Macaque


Dog


Cow


Monodelphis

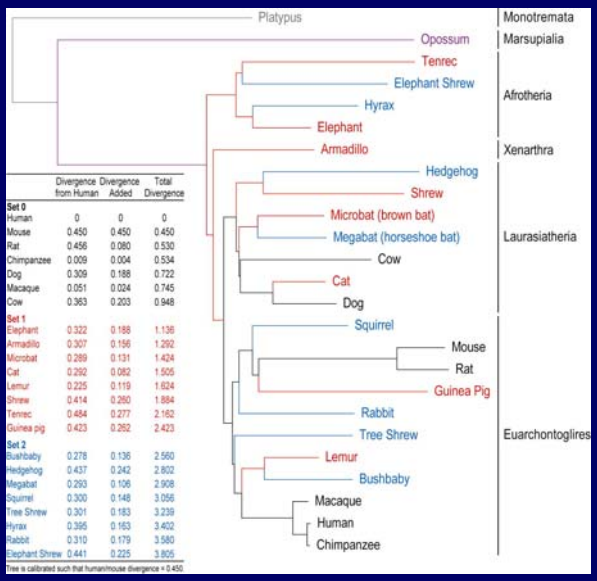
Chicken


Xenopus


Zebrafish


Pufferfish

Low-Redundancy Sequencing of Multiple Vertebrate Genomes



	Divergence from Human	Divergence Added	Total Divergence
Set 0			
Human	0	0	0
Mouse	0.450	0.450	0.450
Rat	0.456	0.080	0.530
Chimpanzee	0.009	0.004	0.534
Dog	0.309	0.188	0.722
Macaque	0.051	0.024	0.745
Cow	0.363	0.203	0.948
Set 1			
Elephant	0.322	0.188	1.136
Armadillo	0.307	0.156	1.292
Microbat	0.289	0.131	1.424
Cat	0.292	0.082	1.505
Lemur	0.225	0.119	1.624
Shrew	0.414	0.260	1.884
Tenrec	0.484	0.277	2.162
Guinea pig	0.423	0.262	2.423
Set 2			
Bushbaby	0.278	0.136	2.560
Hedgehog	0.437	0.242	2.802
Megabat	0.293	0.106	2.908
Squirrel	0.300	0.148	3.056
Tree Shrew	0.301	0.183	3.239
Hyrax	0.395	0.163	3.402
Rabbit	0.310	0.179	3.580
Elephant Shrew	0.441	0.225	3.805

Tree is calibrated such that human/mouse divergence = 0.450

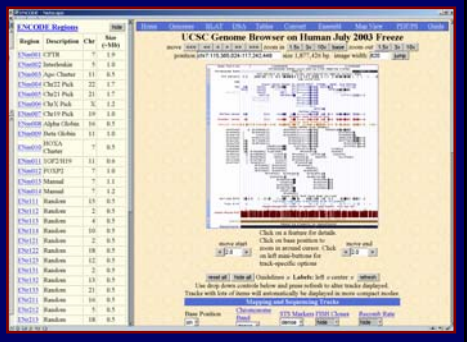
Margulies et al., *PNAS*, in press, 2005

Future Genomes to Sequence???

ENCODE Project

- ENCODE: ENCyclopedia Of DNA Elements
- Goal: Compile a *comprehensive encyclopedia* of all functional elements in the human genome
- Initial pilot project: 1% of human genome
- Apply multiple approaches to study and analyze that 1% in a consortium fashion

ENCODE Project: Web Sites



genome.gov/ENCODE

genome.ucsc.edu/ENCODE

Current Big Challenges…

- Defining "Saturation Points" in Terms of Information Gained by Comparative Sequence Analyses
- The "$1000 Genome"
- Medical Sequencing (aka, Human Re-Sequencing)

Bibliography

Adams, M.D. et al. (2000). The genome sequence of Drosophila melanogaster. Science 287, 2185-2195.

Aparicio, S. et al. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301-1310.

Birren, B. et al. (1998). Bacterial artificial chromosomes. In *Genome Analysis: A Laboratory Manual, Vol. 3 Cloning systems* (B. Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 241-295.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282, 2012-2018.

Collins, F.S. et al. (2003). A vision for the future of genomics research: a blueprint for the genomic era. Nature 422, 835-847.

Rat Genome Sequencing Project Consortium (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493-521.

Goffeau, A. et al. (1997). The Yeast Genome Directory. Nature 387S, 1-105.

Gordon, D. et al. (1998). Consed: a graphical tool for sequence finishing. Genome Research 8, 195-202.

Green, E.D. (2001). Strategies for the systematic sequencing of complex genomes. Nature Rev Genet 2, 573-583.

Green, E.D. (2001). The Human Genome Project and its impact on the study of human disease. In *The Metabolic and Molecular Bases of Inherited Disease, 8th Edition* (C.R. Scriver, A.L. Beaudet, W.S. Sly, D. Valle, B. Childs, K.W. Kinzler, and B. Vogelstein, eds.; McGraw-Hill, Inc.), pp. 259-298.

Green, E.D. et al. (1998). Yeast artificial chromosomes. In *Genome Analysis: A Laboratory Manual, Vol. 3 Cloning systems* (B. Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 297–565.

Gregory, S.G. et al. (1997). Genome mapping by fluorescent fingerprinting. Genome Research 7, 1162-1168.

Hillier, L.W. et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695-716.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.

Jaillon, O. et al. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946-957.

Margulies, E.M. et al. (2005). An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci, in press.

Marra, M.A. et al. (1997). High throughput fingerprint analysis of large-insert clones. Genome Research 7, 1072-1084.

Messing, J. and Llaca, V. (1998). Importance of anchor genomes for any plant genome project. Proc Natl Acad Sci 95, 2017-2020.

Miller, W. et al. (2004). Comparative genomics. Ann Rev Genomics Hum Genet 5, 15-56.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

Shizuya, H. et al. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci 89, 8794-8797.

Thomas, J.W. et al. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424, 788-793.

Venter, J.C. et al. (2001). The sequence of the human genome. Science 291, 1304-1351.

Wilson, R.K. and Mardis, E.R. (1997). Fluorescence-based DNA sequencing. In *Genome Analysis: A Laboratory Manual, Vol. 1 Analyzing DNA* (B. Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 301-395.

Wilson, R.K. and Mardis, E.R. (1997). Shotgun sequencing. In *Genome Analysis: A Laboratory Manual, Vol. 1 Analyzing DNA* (B. Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 397-454.